Seeing the trees in the forest

Diagnosing individual performance in likelihood ratio based forensic voice comparison

Justin J. H. Lo (University of York) @justinjhlo

XVII AISV 2021 - 5 Feb 2021

Evaluating LR-based system performance

Global, system-level metrics

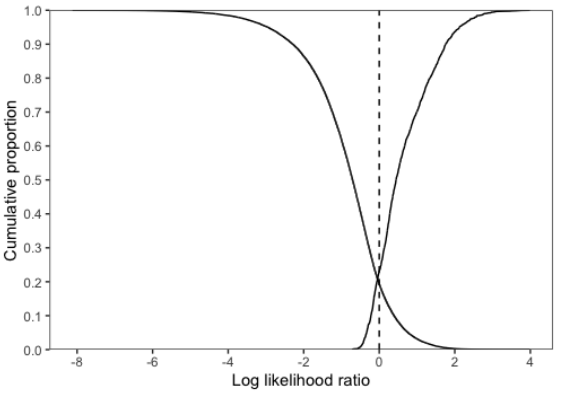
- EER
- C_{llr}

  (Brümmer, et al., 2006)

↓ EER, C_{llr} = ↑ system performance

Graphical means of evaluation

- ROC, DET curves
- Tippett plots ↓



Limitations

- Variation between individual speakers?
- Nature of errors?

» Individual-level analysis

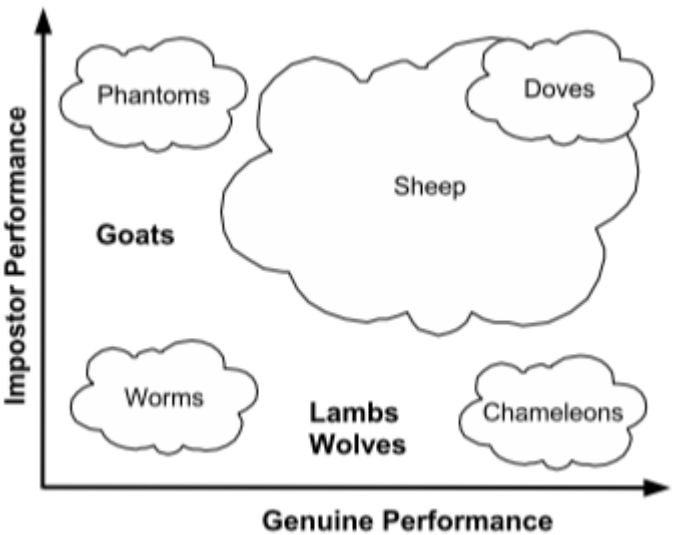
Biometric menagerie

- Classifies speakers into animal groups based on individual performance
  (Doddington, et al., 1998)

- Additional **relational groups** identify subsets of speakers with relatively outlying performance
  - **Dove**: Best-performing (good SS, good DS)
  - **Worm**: Worst-performing (poor SS, poor DS)
  - **Phantom**: Most difficult to match with any speaker (poor SS, good DS)
  - **Chameleon**: Most easily matched with any speaker (good SS, poor DS)

(Dunstone and Yager, 2009)

- Visualised in zooplots
  - SS: stronger →
  - DS: stronger ↑



Biometric menagerie

Zooplot applied to ASR

- Impact of **technical factors** (e.g. SNR) on individual performance
- Potential link between animal groups and **voice quality** (VPA)

(Alexander, et al., 2014; Nash, 2019)

» How effective is zooplot analysis when applied outside ASR?

» Direct connection between animal groups and the input data?

Current study

- Zooplot analysis in LR-based FVC outside ASR context
- Explore connection between individual LR performance and underlying speech data
- Focus on individuals (trees) rather than overall performance (forest)

- Variable: Long-term formant distributions (LTFDs)
  (Nolan and Grigoras, 2005)
  - Collection of formant estimates from all vowels
  - Captures overall articulatory habit and filter behaviour of vocal tract
  - Underlying data readily available for direct comparison

Methodology

Data preparation

- **Materials**: HQ En recordings from 60 male Canadian En-Fr speakers
  (RCMP, 2010-2016)
- **Read speech**: Phonetically balanced short sentences (+ passage)
- **Segmentation**: Automatic forced alignment (manually checked)
  (McAuliffe, et al., 2017)
- **F1-F4 estimates** extracted every 10ms from all vowels (+ /j w/) in Praat
  (Boersma and Weenink, 2016)
- Fixed formant settings (based on preliminary testing):
  - Max 6 formants up to 5500 Hz
  - Window length: 25 ms

LR testing

5 systems tested: individual LTF1-LTF4 + all combined

- Speakers divided into *test* (20), *training* (20) and *background* (20) sets
- LTFDs modelled and compared for all speaker-pairs using **GMM-UBM**
  (Becker, et al., 2008; Reynolds, et al., 2000)
- **Log_{10}LRs** from logistic regression calibration
- 100 repetitions of partitioning and testing
  - Minimise effects of random speaker sampling
  - Ensure all pairs of speakers compared

Individual-level analysis

Zooplots

- Mean LLR in different-speaker (DS) comparisons plotted against mean LLR in same-speaker (SS) comparisons
- From all comparisons across 100 repetitions involving that speaker
- Animal groups defined in accordance with Dunstone and Yager (2009)

DS \ SS	Best 25%	Worst 25%
Best 25%	Doves 🕊️	Phantoms 👻
Worst 25%	Chameleons 🦎	Worms 🪱

- ➕ identify *near*-animals: SS and DS between best/worst 25% and 30%

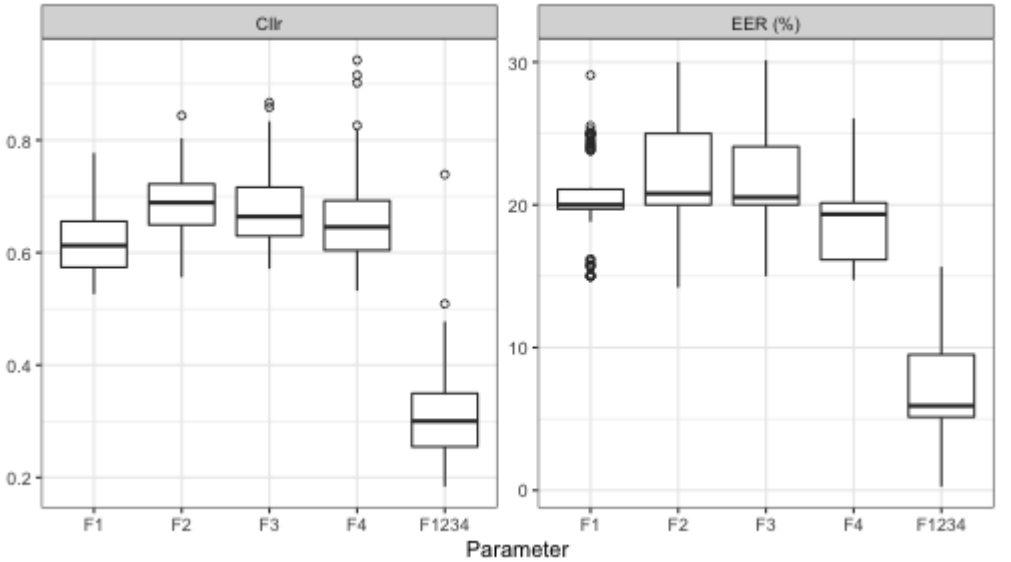
Acoustic comparison for all individual speakers

- Particular focus on members of animal groups

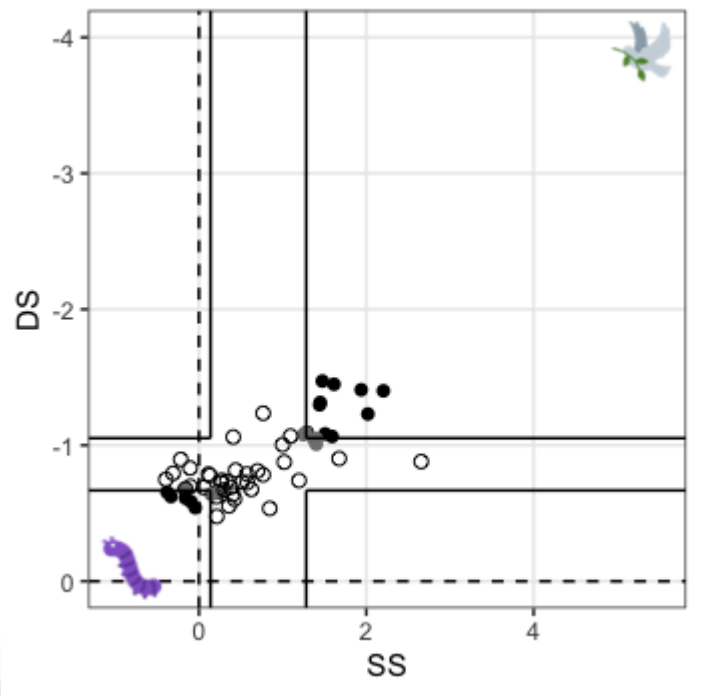
Results & Discussion

Global metrics

- LTF1234 combined performs better than individual LTFs
- Individual LTFs perform on similar levels

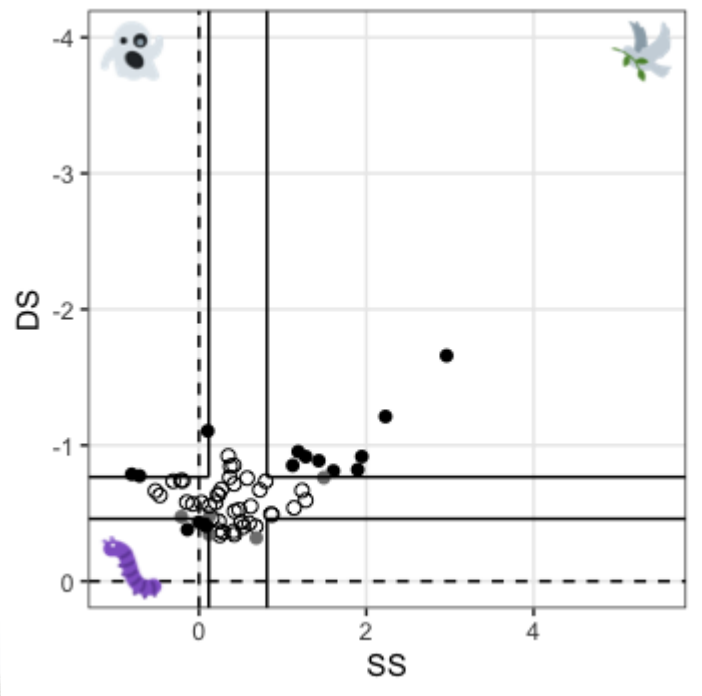


Zooplot: LTF1



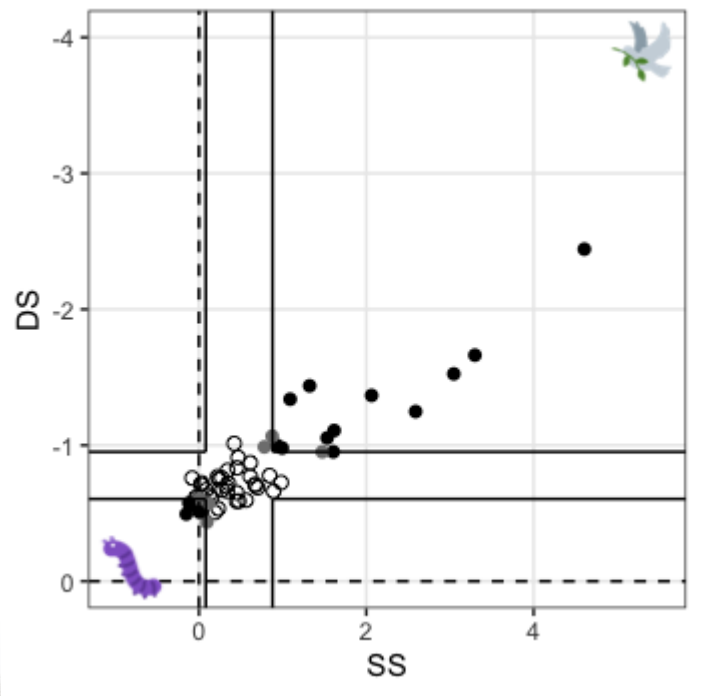
- Narrow ranges of mean LLR
  - SS: -0.4 to 2.66
    DS: -0.48 to -1.47
  - Dense cluster around
    (SS, DS) = (0.5, -0.7)
- SS and DS performance strongly correlated
- 10 doves, 5 worms

Zooplot: LTF2



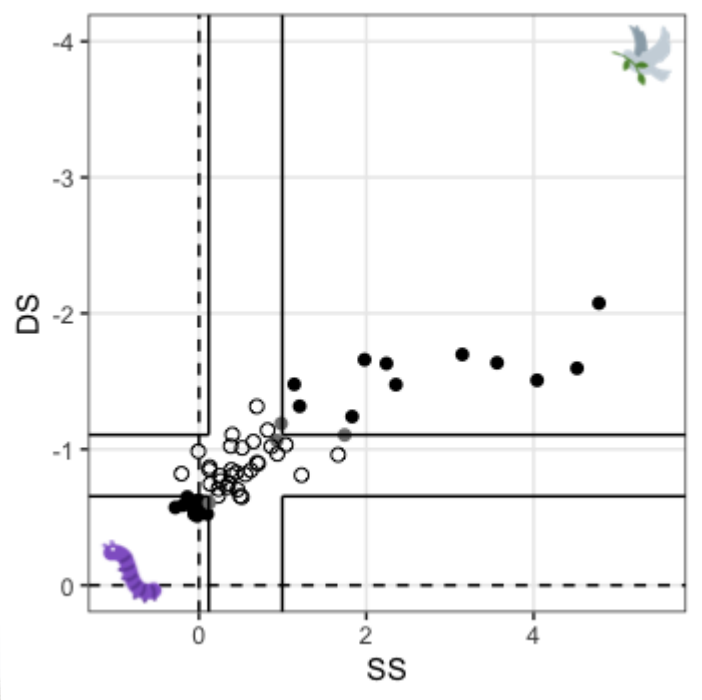
- Relatively poor DS performance
- Negative mean SS-LLRs of greater magnitude
- Weaker correlation between SS and DS performance
- 9 doves, 3 worms, 3 phantoms
- Only system with phantoms

Zooplot: LTF3



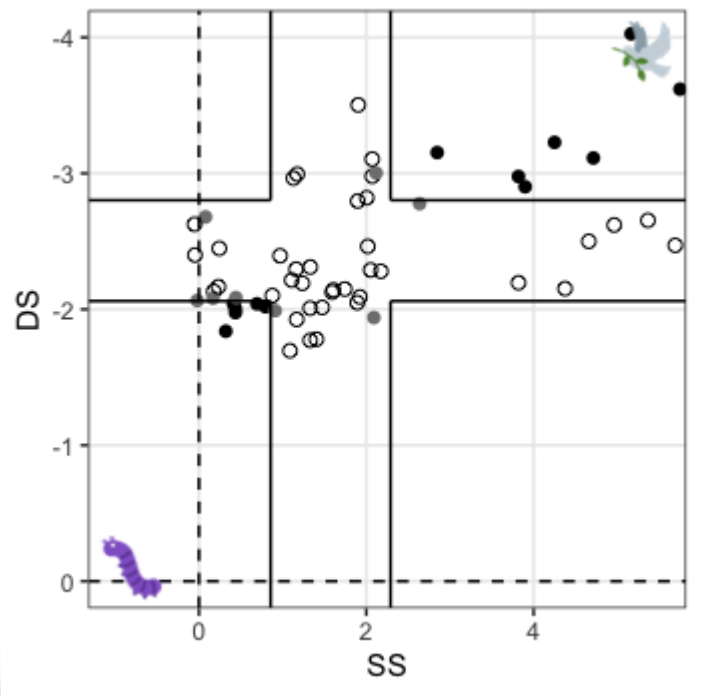
- 12 doves, 8 worms
- Doves with stronger SS and DS performance
- Worms and near-worms clustered near SS = 0

Zooplot: LTF4



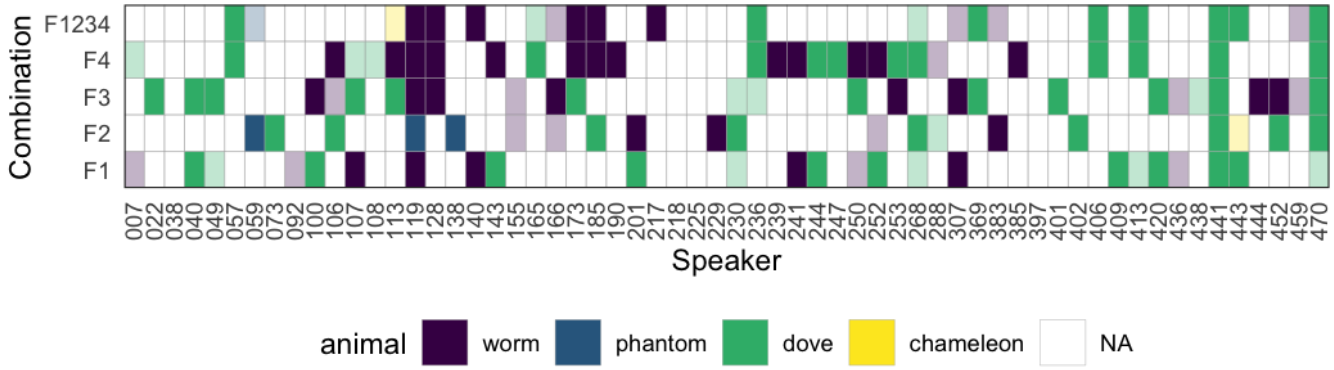
- Overall distribution similar to LTF3 but less clustered
- More extreme distribution: Higher no. of doves and worms
    - 11 doves, 13 worms

Zooplot: LTF1234



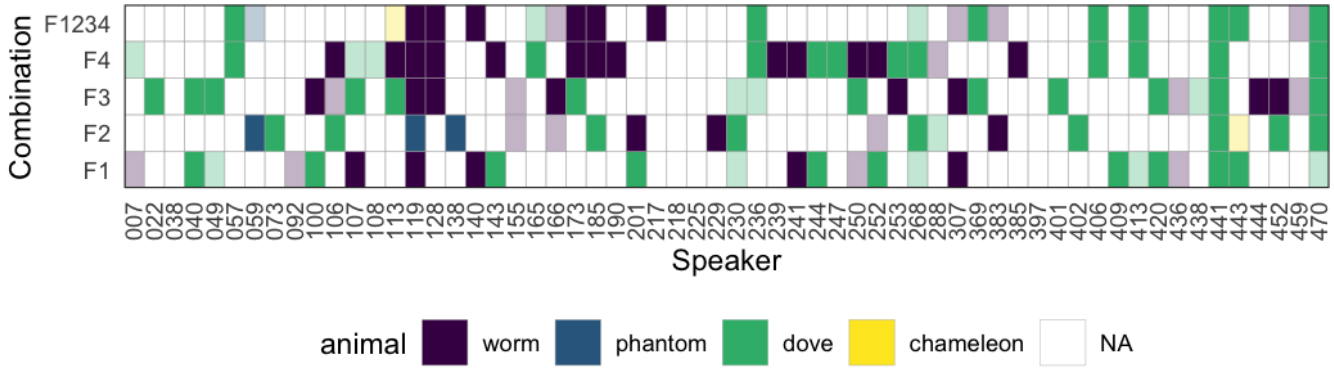
- Better SS and DS performance
- Less clustered distribution: Variation not driven by individual outliers

Zooplot: Summary



- Each LTF captures a very different group of outlying speakers
    - Evidence of complementary speaker-specific information on individual level

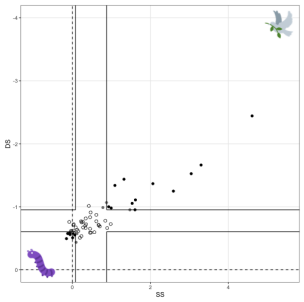
Zooplot: Summary

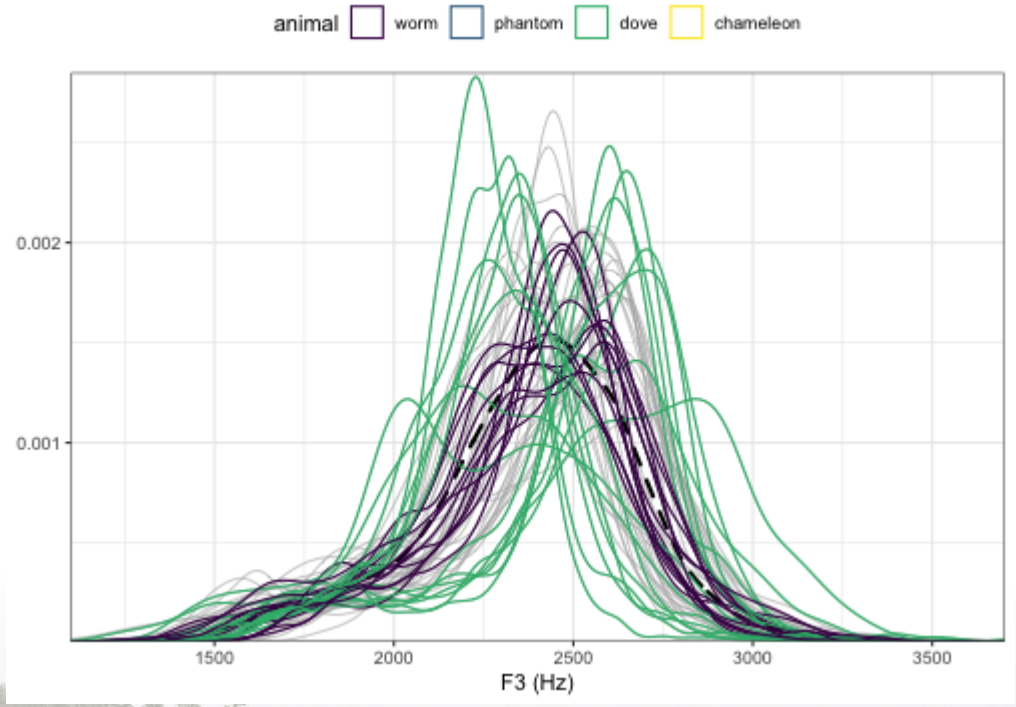


- Across all LTFs:
  - Only 3 speakers (5%) **always** in/near groups: `119`, `441`, `470`
  - 5 (8%) speakers **not** in/near any group: `038`, `217`, `218`, `225`, `397`

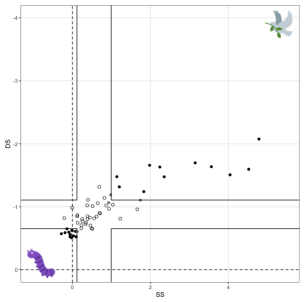
Acoustic comparisons

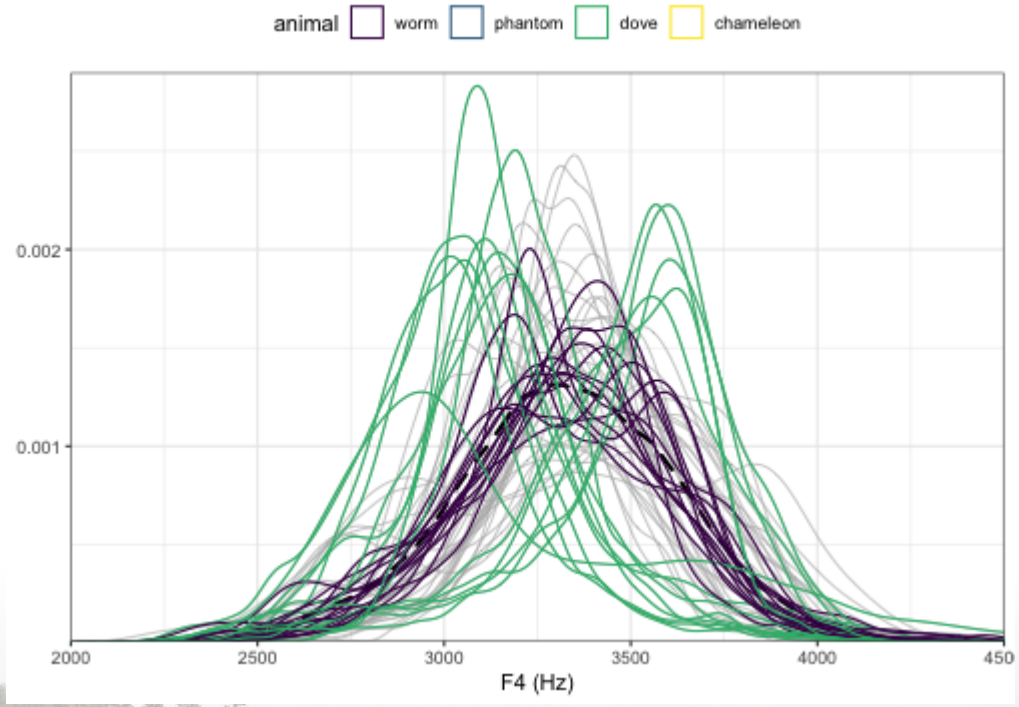
F3



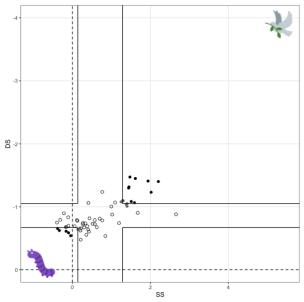

animal
worm
phantom
dove
chameleon
0.002
0.001
1500
2000
2500
3000
3500
F3 (Hz)

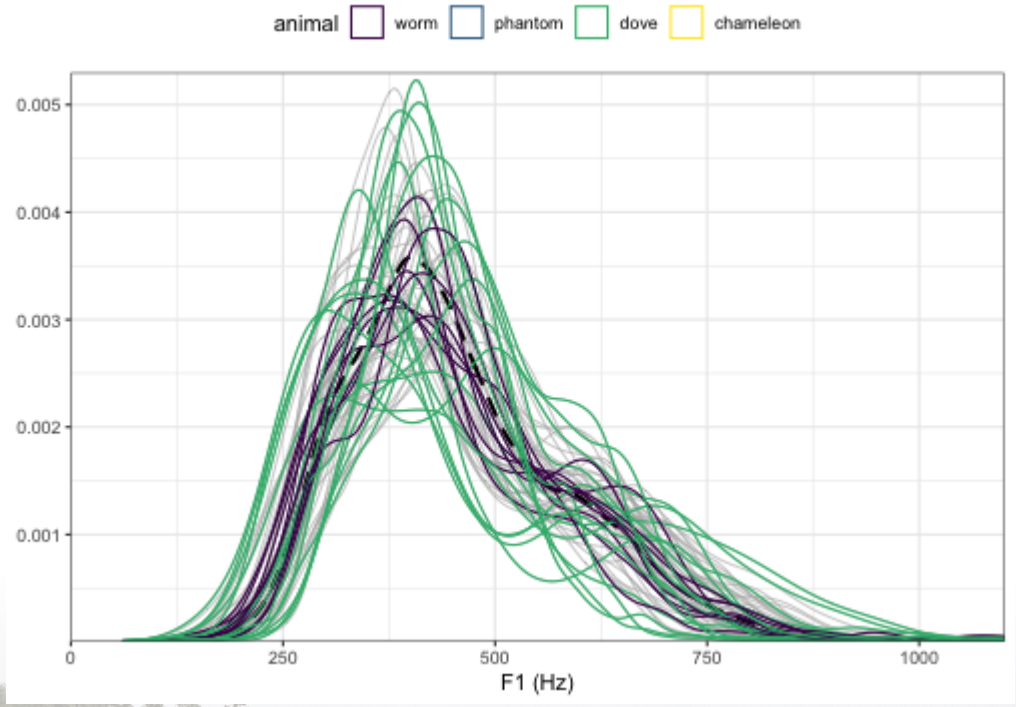
F4



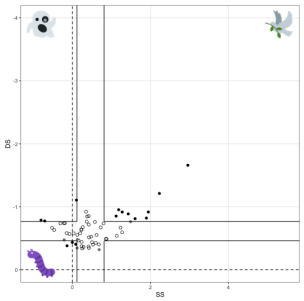

animal
worm
phantom
dove
chameleon
0.002
0.001
2000
2500
3000
3500
4000
450
F4 (Hz)

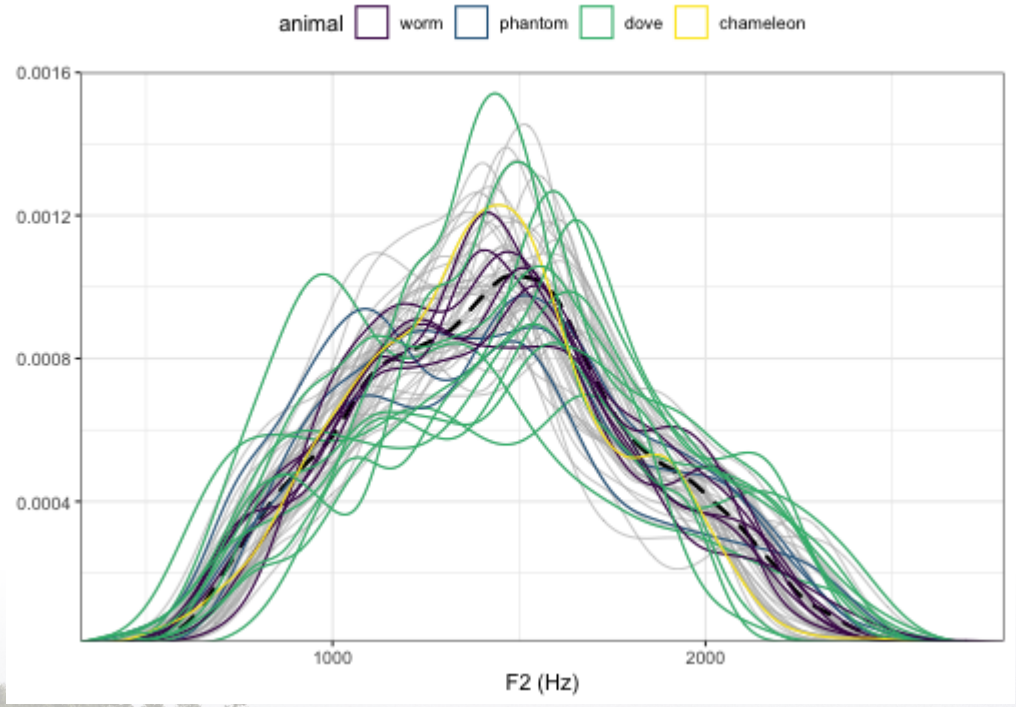
F1




animal
worm
phantom
dove
chameleon
0.005
0.004
0.003
0.002
0.001
0
250
500
750
1000
F1 (Hz)

F2




animal
worm
phantom
dove
chameleon
0.0016
0.0012
0.0008
0.0004
1000
2000
F2 (Hz)

Thank you!

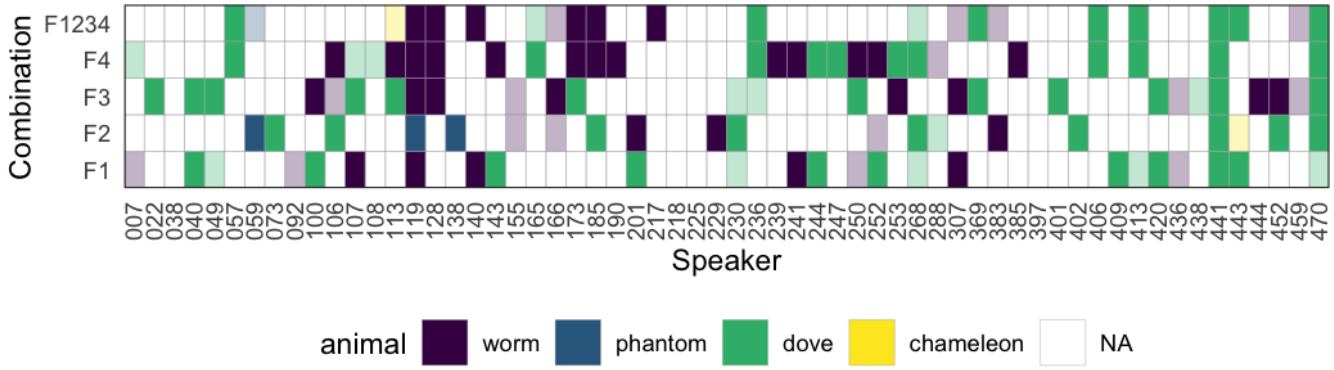
Questions?

@justinjhlo JL2355@york.ac.uk

Many thanks to:

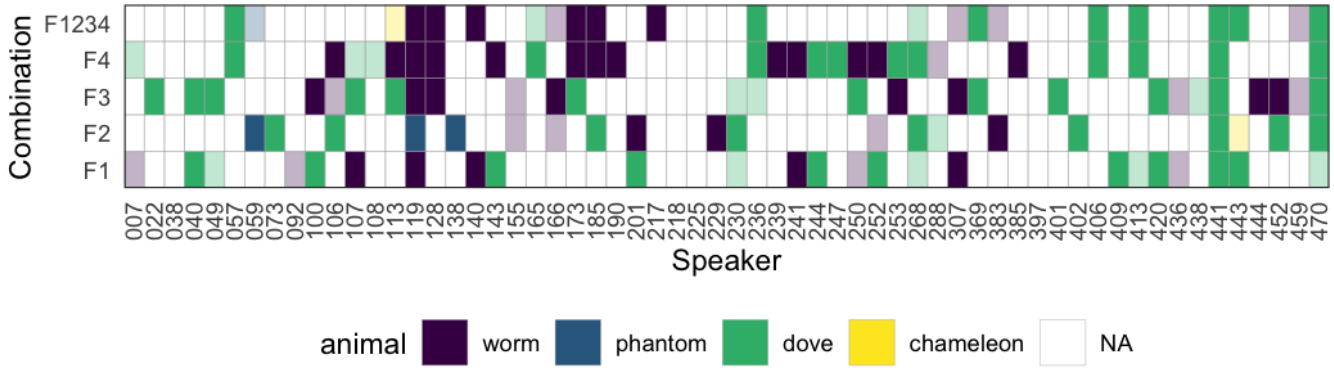
- Paul Foulkes and Vincent Hughes for guidance and feedback
- Royal Canadian Mounted Police for use of the *Voice ID Database*

Zooplot: Summary



- 21 speakers (35%) in/near the same group for more than one F *vs* 13 speakers (22%) in/near both doves and worms at least once

Zooplot: Summary



- All speakers in/near a group for LTF1234 also in/near same group for at least one of LTF1-LTF4
  - Most in common with LTF4
  - Exception: 217