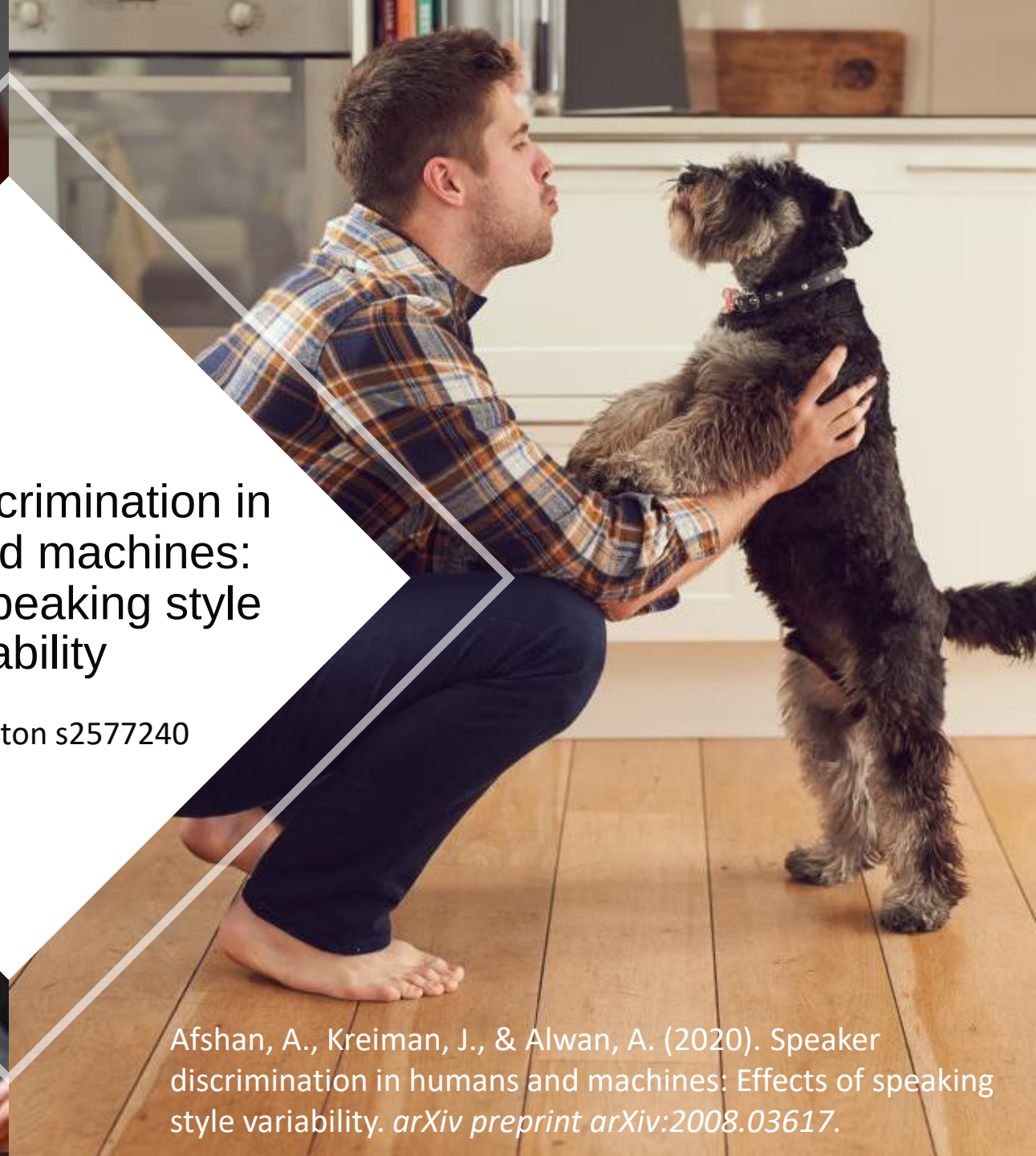
Speaker discrimination in humans and machines: Effects of speaking style variability

Chris Congleton s2577240



Afshan, A., Kreiman, J., & Alwan, A. (2020). Speaker discrimination in humans and machines: Effects of speaking style variability. *arXiv preprint arXiv:2008.03617*.

Introduction

- Speaker discrimination
  - Recognising speaker
- Speaking style variability
  - Talking to a friend
  - Reading aloud
  - Calling a pet

Research Questions

- Unfamiliar speech recognition task
  - Two samples, same speaker?
  - Read and conversational
- Humans vs Computers
  - Change in speaking style
  - Natives listener performance

Data

- UCLA Speaker Variability Database [1]
  - Female only
  - Read → Phonetically rich Harvard sentence
  - Conversation → Telephone family/friend
- Automatic Speaker Verification system [2]
  - NIST Speaker Recognition Evaluation database
  - 3,000 hours of speech samples

1. P. Keating, J. Kreiman, and A. Alwan, “A New Speech DatabaseFor Within- and Between-Speaker Variability,”Proc of the 19thICPhS, p. 4, 2019.
2. M. Przybocki and A. Martin, “NIST Speaker Recognition Evaluation Chronicles,” inProc. Odyssey, 2004, pp. 12–22.

Methods

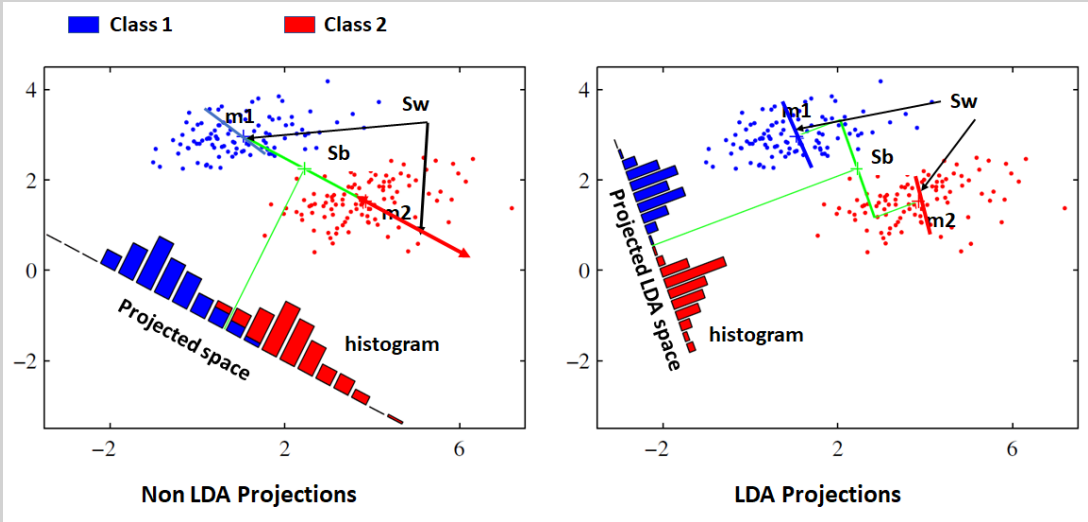
- Tasks: Determine same speaker?
  - Two read sentences
  - Two conversational sentences
  - One read, one conversational
- Constant ratio
  - Same speaker (target)
  - Different speaker (non-target)
- Samples
  - 3 seconds
  - All distinct
  - Non-speech vocalisations removed

Human Discriminator

- Played only once
  - Same or different speaker
  - Confidence 0 – 5
  - About 45 min of testing
  - At own pace

Automatic Speaker Verification (Computer discriminator)

- Probabilistic Linear Discriminant Analysis
  - Dimensionality reduction
  - Maximising spread between class means
  - Minimising spread of class
  - Handles unseen class
- x-vector embeddings
  - Time Delay Neural Network
  - Variable length utterences
  - Mel-frequency cepstral coefficients as features



Source: *"Pattern Recognition and Machine Learning" by Bishop. 2006*

Evaluation

- Continuous similarity score
  - Decision: same = 1, different = -1
  - Confidence 0 – 5
  - From -5 up to 5
- PLDA score
  - Probability of same/different
- Equal error rate
  - Intersection of False Acceptence and False Rejection
  - Lower is more accurate

Results

- Style matched
- Nativity
- System fusion

Listener	Read-read	Conversation-conversation	Read-conversation
Native	EER%	EER%	EER%
Machines	14,35	19,87	21,78
Humans	6,96	15,12	20,68
Fusion	4,92	11,20	16,39
Non-native			
Machines	13,95	19,47	19,64
Humans	12,39	23,22	31,46
Fusion	5,69	13,57	19,34

Conclusion

- System fusion
  - Different discrimination strategies
- Style match improves performance
  - Read vs conversation
  - Variability acoustic spaces
- Humans more confident for same speaker tasks
- Difficulty of speaker different for each system