Classifying non-speech vocalisations for speaker recognition

Finnian Kelly[1], Harry Swanson[2], Kirsty McDougall[2], and Anil Alexander[1]
[1]*Oxford Wave Research Ltd., Oxford, U.K.*
[2]*University of Cambridge, Cambridge, U.K.*

IAFPA 2022 conference: 11/07/22

Non-speech vocalisations (NSVs)

NSVs are sounds speakers can produce with their vocal organs that do not have linguistic content, and may or may not contribute meaning to a communication

- Examples of NSVs: laughs, screams, roars, yawns, moans, groans, sighs, coughs, throat-clearings, hiccups, sneezes, paralinguistic clicks
- NSVs broadly fall into two groups:
  - Auditory reflexes of physiological processes, e.g. non-volitional coughs, yawns, throat-clearing
  - Extralinguistic calls of an emotional nature, e.g. laughs, screams, groans, moans

Non-speech vocalisations (NSVs)

NSVs are sounds speakers can produce with their vocal organs that do not have linguistic content, and may or may not contribute meaning to a communication

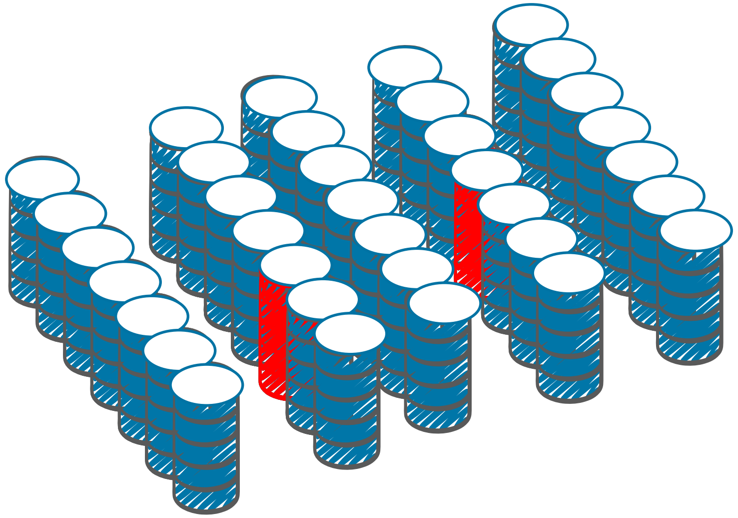
- Examples of NSVs: laughs, screams, roars, yawns, moans, groans, sighs, coughs, throat-clearings, hiccups, sneezes, paralinguistic clicks
- NSVs broadly fall into two groups:
  - Auditory reflexes of physiological processes, e.g. non-volitional coughs, yawns, throat-clearing
  - Extralinguistic calls of an emotional nature, e.g. laughs, screams, groans, moans

Not often used in forensic speaker recognition but may contain speaker-characterising information

NSVs can be important

Research Question 1: *Investigative*

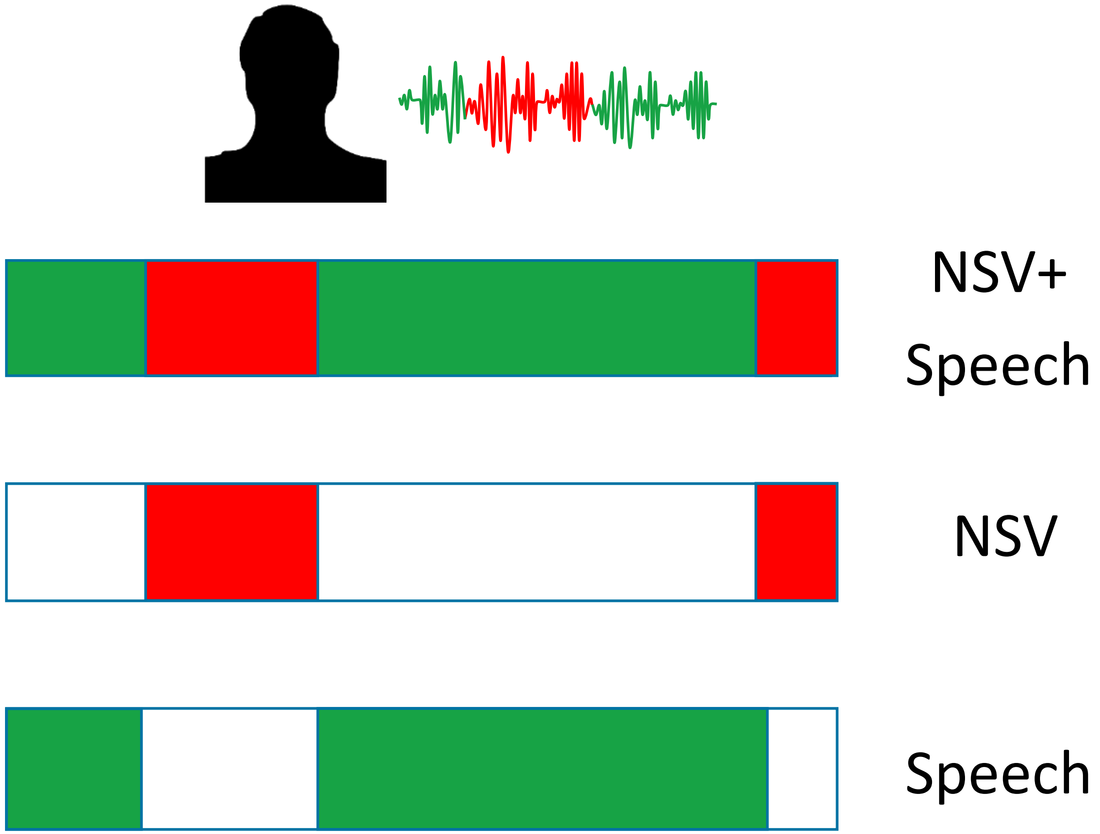
- In a large set of speech recordings, can we find those containing NSVs of interest?
  - Can we distinguish between specific types of NSVs (e.g., screams, moans, laughs)?
  - Can we find the location of the NSV in the recording?



Example scenario: triage of a large dataset to find those audio/video recordings containing screams or moans

Research Question 2: *Forensic*

- How do NSVs affect automatic speaker recognition?
  - If a recording contains an NSV and speech, is it better to remove or preserve the NSV?
  - If a recording contains an NSV and no speech, can automatic speaker recognition still be applied?



Example scenario: comparison of a known voice with questioned recordings containing screams or moans and only sparse amounts of speech

NSVs and speaker recognition: what we know

- NSVs are typically discarded prior to automatic speaker recognition modelling and comparison
- Research involving NSVs and speaker recognition is limited, but there are some findings that show certain NSVs contain *speaker-characterising information:*
  - Human listeners: above-chance recognition of speakers based on **Laughs** (Philippon et al., 2013), **Screams** (Engelberg et al. 2019), and **Cries** (Gustafson et al. 1984).
  - Automatic: above-chance recognition of speakers based on **Laughs** (Bacharowski et al., 2001), and **Screams** (Hansen et al., 2017).
- Much of the existing research is based on the comparison of NSVs only; however, the comparison of NSVs with speech is of particular relevance for forensic and investigative speaker recognition

Naturally-elicited NSV data: *Anikin & Persson corpus*

- The corpus contains audio recordings of 603 naturally-elicited NSVs, each produced in a single emotional state by a unique individual
- Audio extracted from YouTube videos, and the video context was used to determine the emotion of a vocalisation (e.g. retching in disgust while unblocking a toilet)
- Each clip is labelled with one of nine **emotional categories** (*amusement, anger, disgust, effort, fear, joy, pain, pleasure, sadness*) and one of eight **call types** (*grunt, laugh, moan, roar, scream, sigh, tone, whimper*)
- Each call type (i.e. NSV) can <u>encode multiple emotions</u>
- Anikin & Persson initially focused on recognition of emotional categories by human listeners
- Subsequently they found that call types (NSVs) may be a more natural categorisation for listeners (Anikin, Bååth & Persson, 2018)

Anikin, A., & Persson, T. (2017). Nonlinguistic vocalizations from online amateur videos for emotion research: A validated corpus. Behavior research methods, 49(2), 758-771.

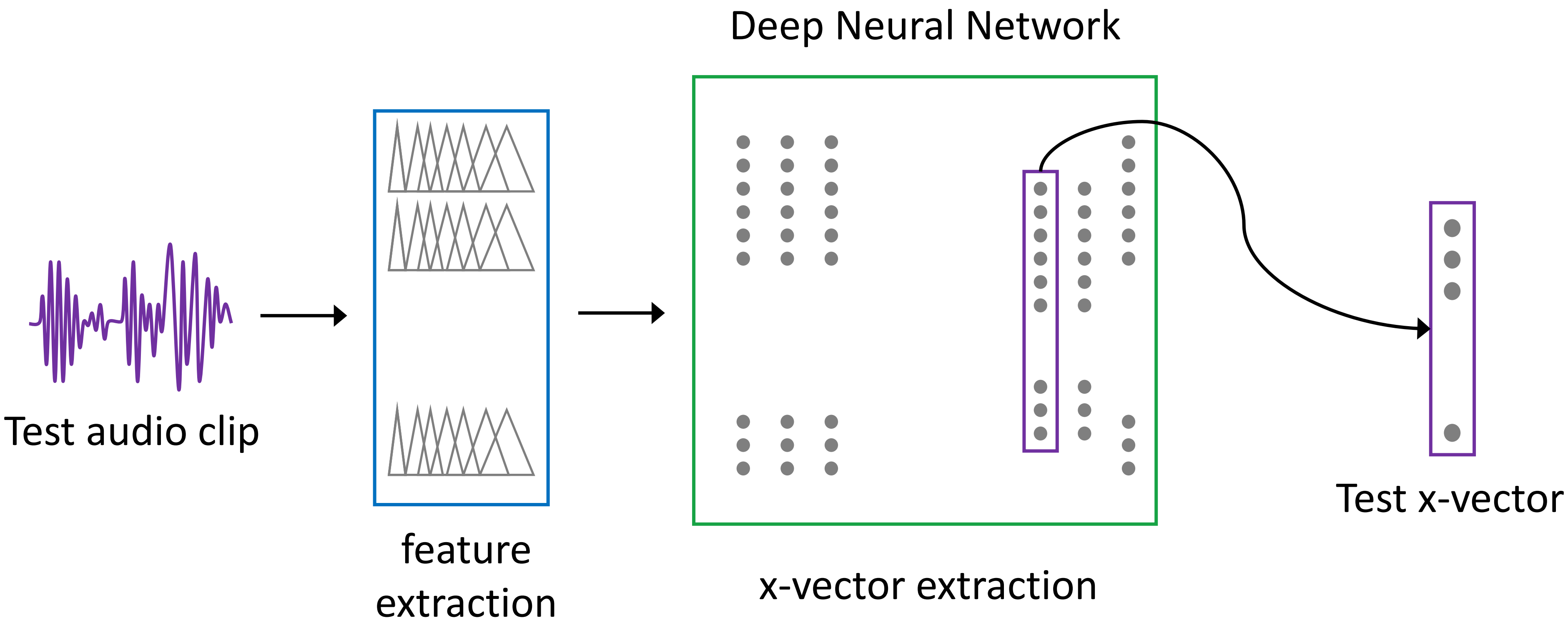
Naturally-elicited NSV data: *Anikin & Persson corpus*

- A subset of four call types were selected as NSVs for our experiments
  - scream (N=91)
  - roar (N=84)
  - laugh (N=109)
  - moan (N=38)

- Additionally, a speech category (N=100) was created by extracting short audio clips of spontaneous speech from YouTube videos (VoxCeleb dataset):

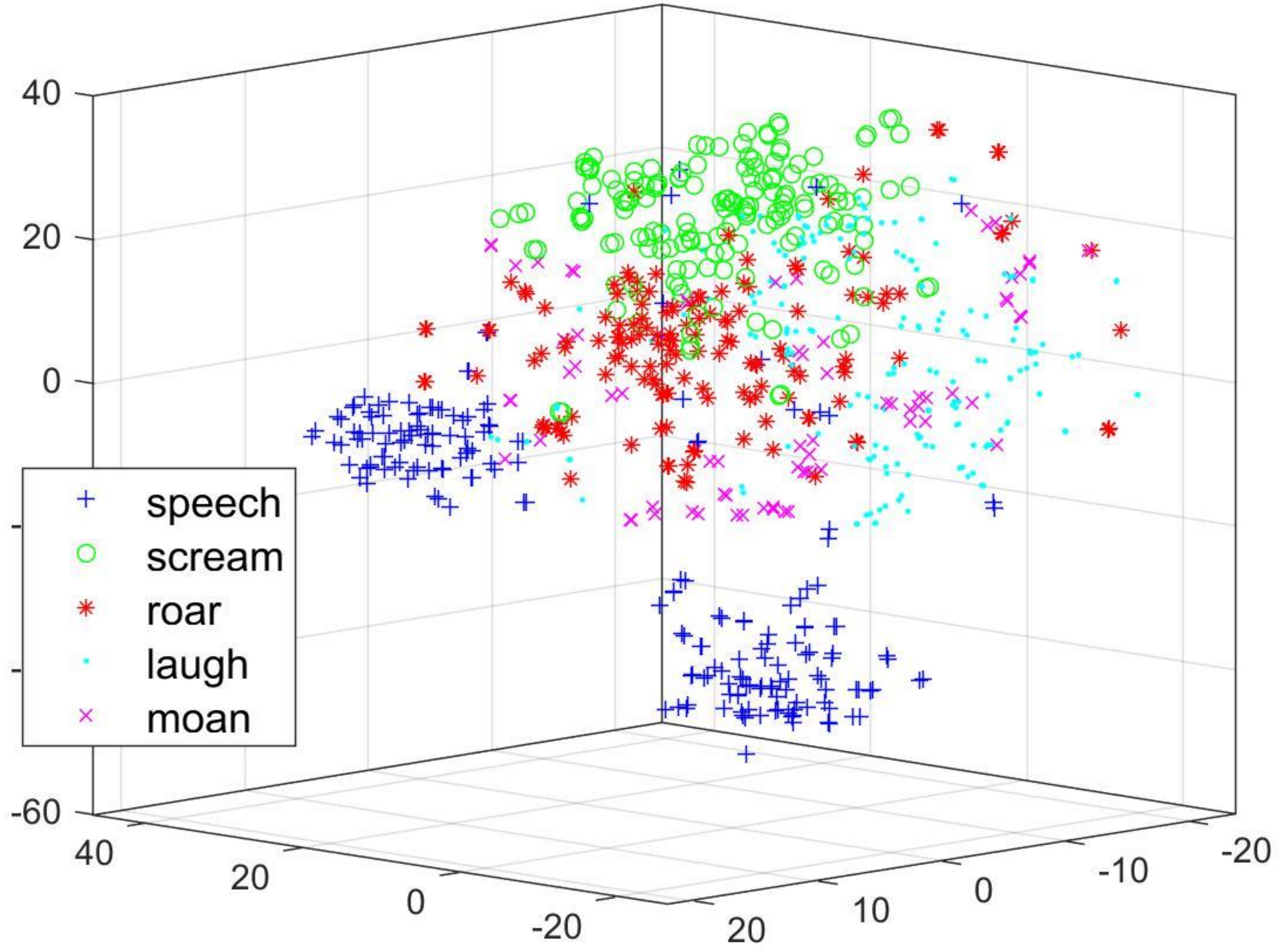
- The NSV recordings are short: 0.5 – 13.6 s (median = 1.7 s). The speech category recordings were trimmed to a similar duration distribution.

NSV classification step 1: x-vector extraction



Kelly, F., Forth, O., Kent, S., Gerlach, L., & Alexander, A. (2019). Deep Neural Network Based Forensic Automatic Speaker Recognition in VOCALISE using x-Vectors, 2019 AES International Conference on Audio Forensics.

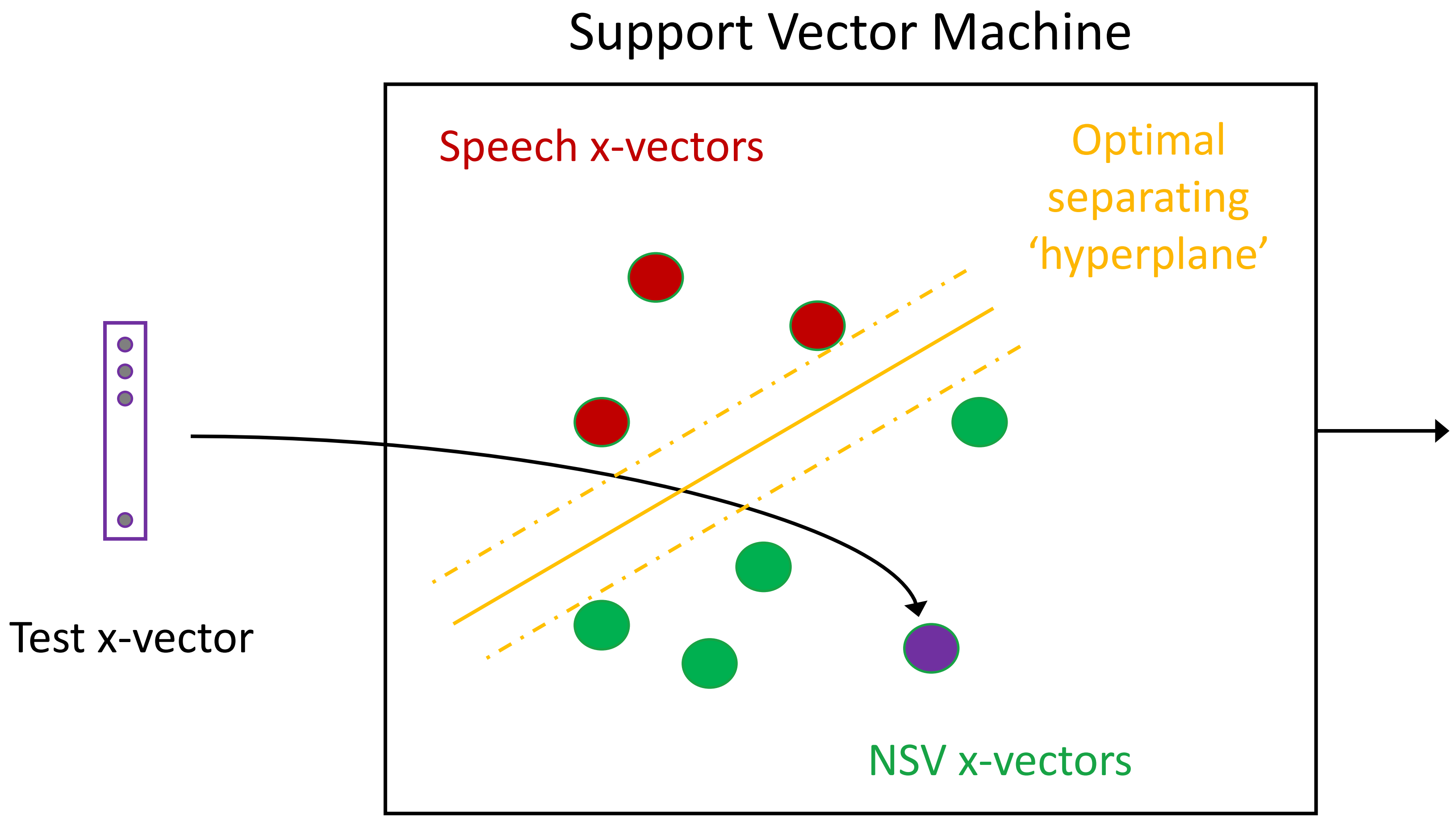
Visualising NSV x-vectors



x-vectors projected into three dimensions using tSNE

The two speech clusters correspond to male and female speakers

NSV classification step 2: x-vector classification

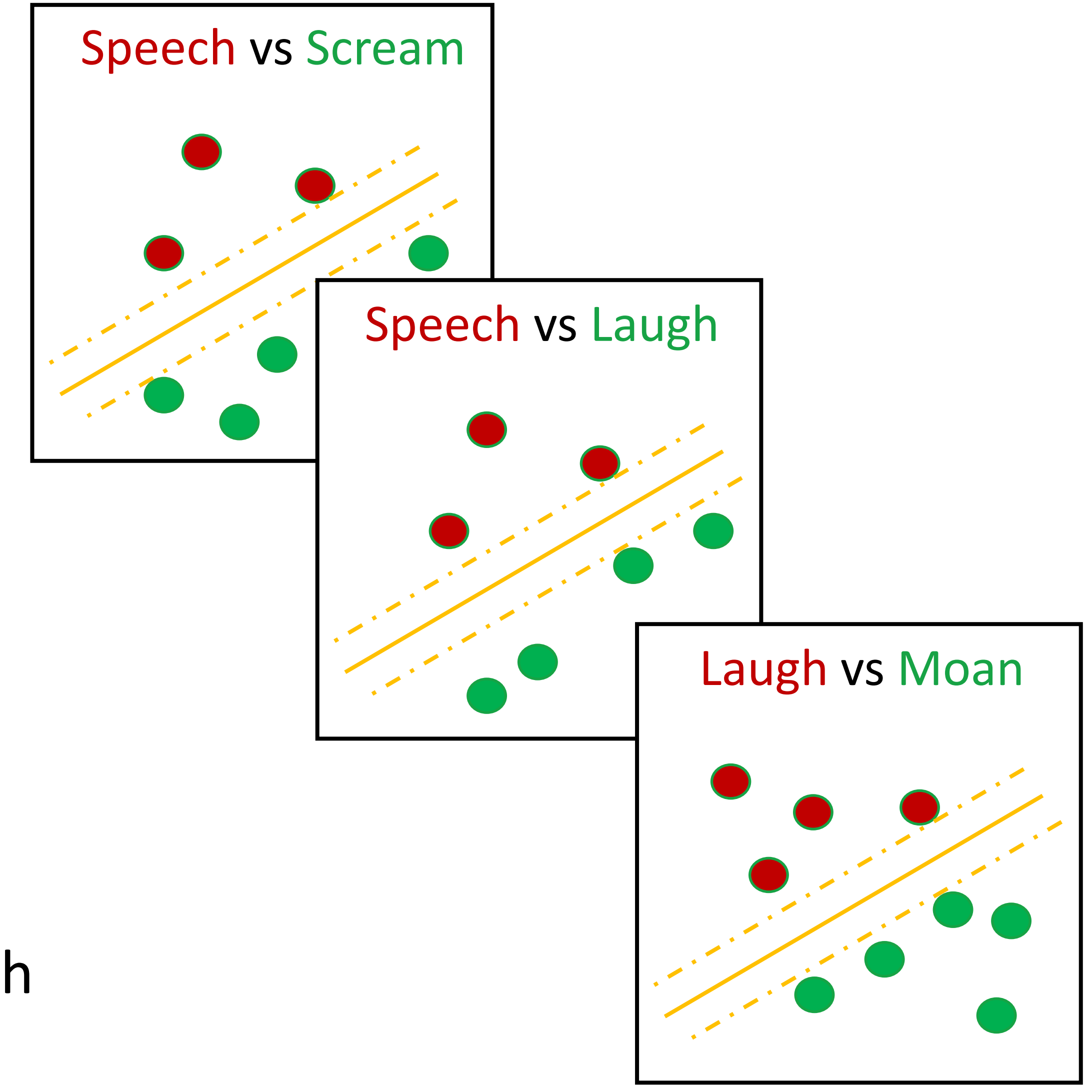


Output label = *NSV*

Output score = *distance from hyperplane*

NSV two-class classification experiment

- An SVM classifier was trained and tested for all two-class combinations of the 5 classes (4 NSV, 1 speech), i.e.,
  - Speech vs Scream
  - Speech vs Laugh
  - ⋮
  - Laugh vs Moan
- For each combination, recordings were split into training and testing sets in the ratio 3:1
- This process was repeated 10 times, each with a different random split



NSV two-class classification experiment results

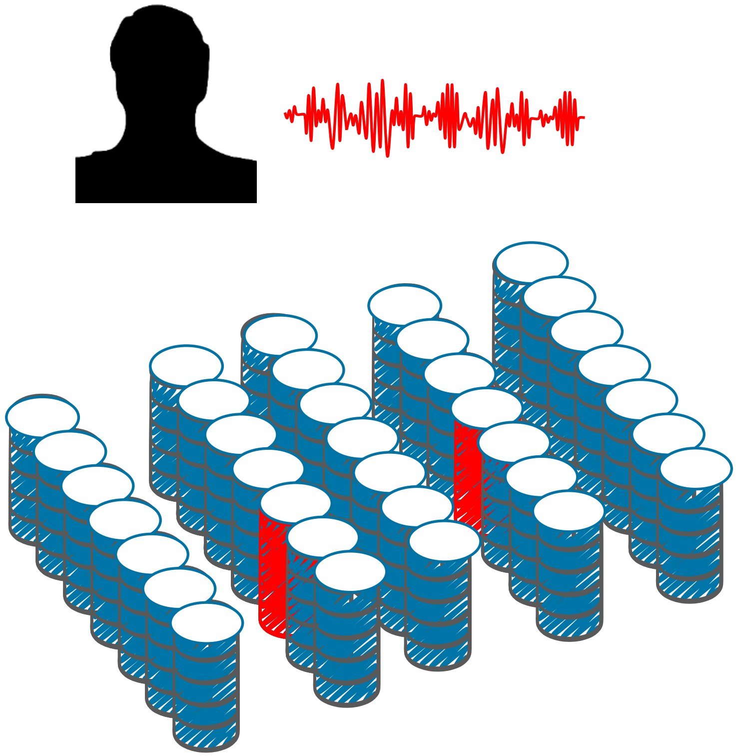
	Speech	Scream	Roar	Laugh	Moan
Speech		0	0.5	0.7	0.3
Scream			11	7.5	7.2
Roar				5.6	9.3
Laugh					9.6

Classification Equal Error Rates (EERs) %

- For Speech vs NSVs, all EERs are <1%
- Most-confusable NSVs are Scream and Roar (11%)
- Least-confusable NSVs are Laugh and Roar (5.6%)

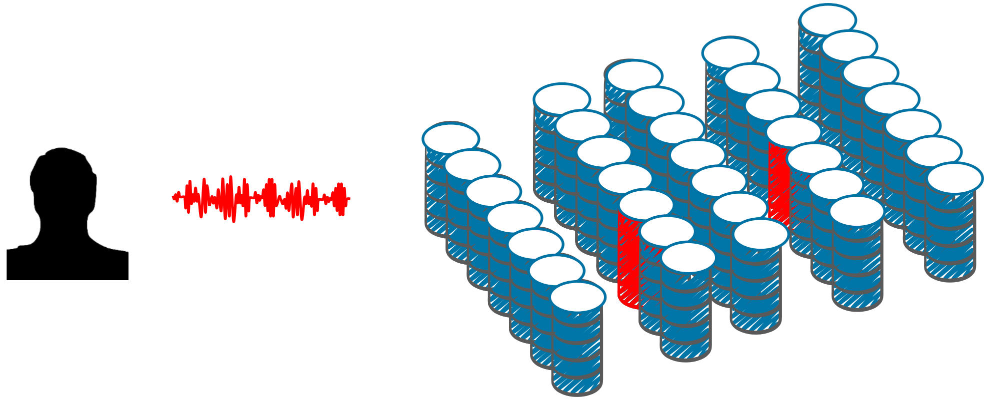
Investigative: proof of concept

This proof-of-concept classification experiment demonstrated that NSVs can be reliably distinguished from speech using an automatic approach, and that different NSVs can be distinguished from each other

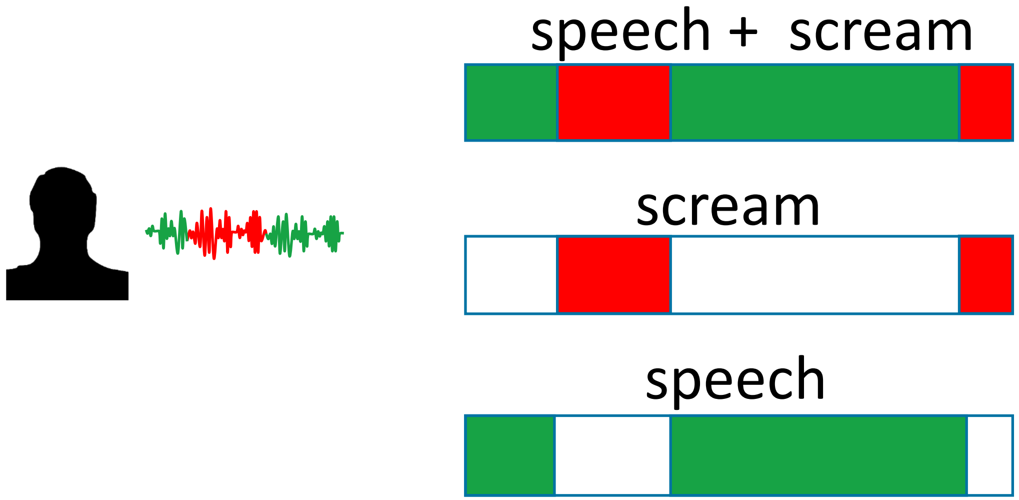


Revisiting a scenario from our research questions

1. *Investigative:* Can we triage a large dataset to find those audio/video recordings containing **screams**?

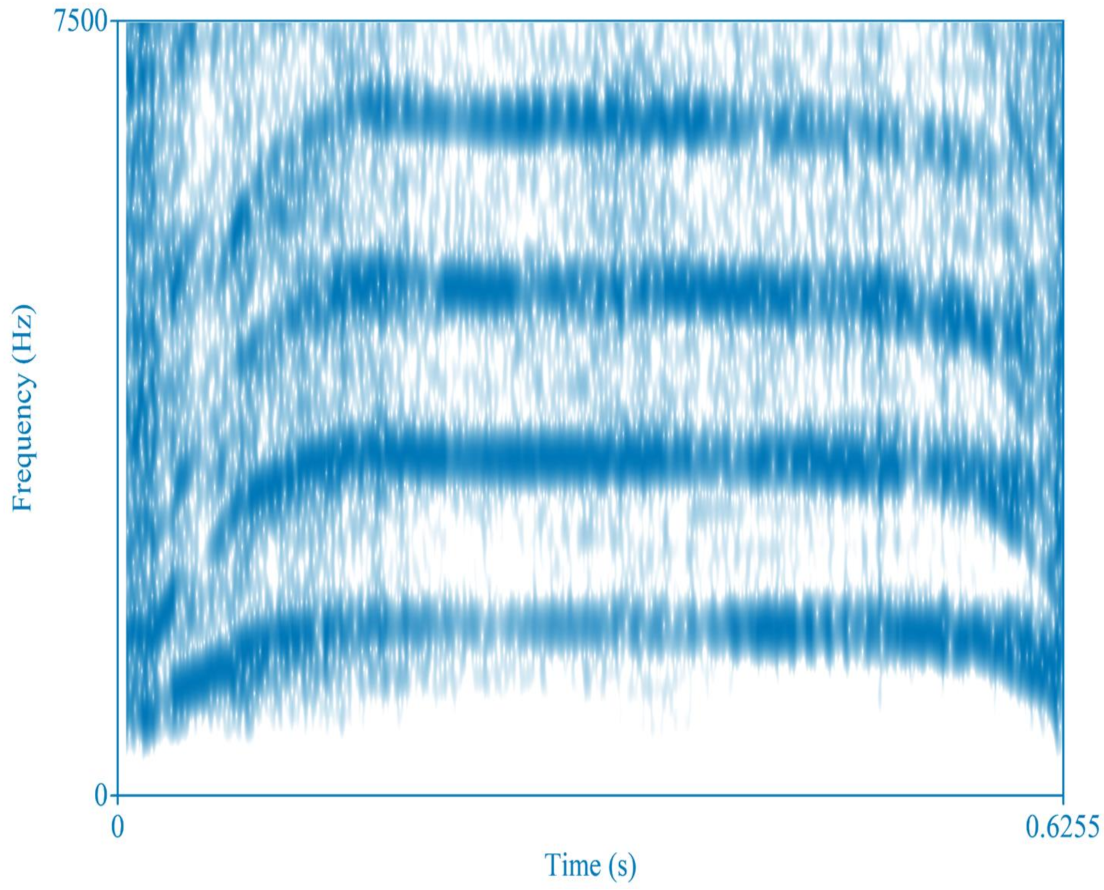


2. *Forensic:* Can we use automatic speaker recognition to compare a known voice with a questioned recordings containing **screams** and only sparse amounts of speech?



What are screams?

- Generally recognised as a loud, high-pitched, usually sustained non-speech vocalisations of high emotional intensity
- Associated with various emotions/states - most commonly fear, followed by pain, excitement/surprise, anger
- Characterised acoustically by:
  - High fundamental frequency
  - High intensity
  - Relatively high formant frequencies (especially F1) due to tongue retraction
  - Relatively uniform energy distribution across frequencies, compared with speech
  - Low number of discrete vocal bursts



F0 = c. 1500Hz
Intensity = c. 80 dB

Naturally-elicited scream data

- The Anikin & Persson NSV corpus has no speech content or speaker labels
- The Speakers in the Wild (SITW) database (McLaren et al. 2016) was therefore used as a source of both naturally-elicited scream data and spontaneous speech
- SITW contains diverse speech content, including 'ice bucket challenge' recordings, many of which contain both speech and screams from the same speaker.

A test set of 'ice bucket challenge' recordings was created by selecting those with only one speaker, and discarding those with very high noise levels (< 5 dB SNR) or very little speech (< 5 sec net). The resulting test set contained:

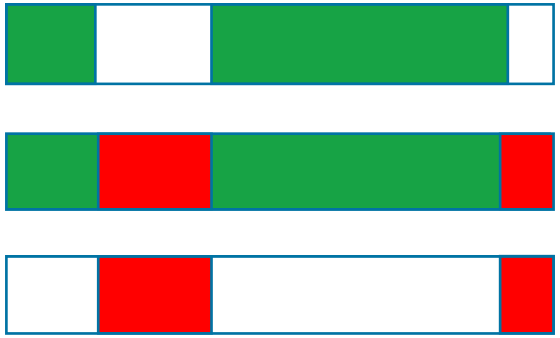
- 20 recordings with speech and scream, each from a unique speaker
- 20 additional speech recordings, one for each of the same 20 speakers

Detecting screams 'in the wild'

- The x-vector SVM Speech vs Scream classifier was retrained incorporating data augmentation for improved performance in noise
- The classifier was applied to short chunks (1 second net, 50% overlap) of the 20 SITW speech-and-scream recordings, and the 20 SITW speech-only recordings
- The maximum chunk score per-recording was selected, and if above 0.5, the recording was labelled as containing scream:
  - 19/20 speech-and-scream recordings were labelled correctly
    - In all cases, the maximum scoring chunk correctly located the scream within the recording
  - 20/20 speech-only recordings were labelled correctly

Speaker recognition with screams "in the wild"

- Given the SITW speech and scream recordings, three conditions were considered:
  1. **Speech**: 5 sec net speech
  2. **Speech-and-scream**: 5 sec net speech + all available scream (0.5-2.5 sec)
  3. **Scream**: all available scream (0.5-2.5 sec)



- For each condition, the SITW speech-only recordings were used as a comparison set in a speaker recognition test*:
  1. **Speech** vs speech-only = 8.7% EER
     - Allowing maximum speech duration (median 7 sec.) = 6.7% EER
  2. **Speech-and-scream** vs speech-only = 11.2% EER
  3. **Scream** vs speech-only = 44.4% EER

*VOCALISE 2021 with an R&D x-vector session

Can NSVs be used reliably for speaker recognition?

Conclusions: *investigative*

- Using an automatic approach, it is possible to:
  - Reliably distinguish NSVs from speech
  - Accurately Locate NSVs (screams) within a larger recording of speech

- Considerations:
  - Distinguishing between different types of NSVs is more challenging
  - Background noises (e.g. car engines, strong wind) may lead to false alarms
  - Very animated/emotional speech may lead to false alarms

Conclusions: *forensic*

- Screams do not benefit automatic speaker recognition
  - Holding speech duration constant, performance decreased with the addition of scream
  - Comparing speech to scream resulted in very poor performance (just above chance)
  - Our findings align closely with those of Hansen et al., 2017

- Considerations:
  - A small sample size was used here, and the screams were short
  - Did not have 2+ screams per speaker to evaluate scream vs scream recognition
  - As the ratio of net speech to scream in a recording increases, the presence of the scream will become less important

Hansen, J. H., Nandwana, M. K., & Shokouhi, N. (2017). Analysis of human scream and its impact on text-independent speaker verification. *The Journal of the Acoustical Society of America, 141*(4), 2957-2967.

References

- Anikin, A., & Persson, T. (2017). Nonlinguistic vocalizations from online amateur videos for emotion research: A validated corpus. Behavior research methods, 49(2), 758-771.
- Anikin, A., Bååth, R., & Persson, T. (2018). Human non-linguistic vocal repertoire: call types and their meaning. Journal of nonverbal behavior, 42(1), 53-80.
- Bachorowski, J. A., M. J. Smoski and M. J. Owren (2001). The acoustic features of human laughter. *Journal of the Acoustical Society of America* 110(3): 1581-1597.
- Engelberg, J. W., J. W. Schwartz and H. Gouzoules (2019). Do human screams permit individual recognition? *PeerJ, 7*, e7087.
- Gustafson, G. E., J. A. Green and T. Tomic (1984). Acoustic correlates of individuality in the cries of human infants. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology* 17(3): 311-324.
- Hansen, J. H., Nandwana, M. K., & Shokouhi, N. (2017). Analysis of human scream and its impact on text-independent speaker verification. The Journal of the Acoustical Society of America, 141(4), 2957-2967.
- Kelly, F., Forth, O., Kent, S., Gerlach, L., & Alexander, A. (2019). Deep Neural Network Based Forensic Automatic Speaker Recognition in VOCALISE using x-Vectors, 2019 AES International Conference on Audio Forensics.
- Philippon, A. C., L. M. Randall and J. Cherryman (2013) "The impact of laughter in earwitness identification performance." Psychiatry, Psychology and Law 20(6): 887-898.
- McLaren, M., Ferrer, L., Castan, D., & Lawson, A. The “Speakers in the Wild (SITW) Speaker Recognition Database.” Interspeech, San Francisco, Sept., 2016.

Thank you for listening!