Feature Representations for Automatic Meerkat Vocalization Classification

Imen Ben Mahmoud[1], Eklavya Sarkar[1,2], Marta Manser[3], Mathew Magimai Doss[1]

[1] Idiap Research Institute, Switzerland
[2] École polytechnique fédérale de Lausanne, Switzerland
[3] University of Zurich, Switzerland

VIHAR Interspeech 2024

September 2024

Contents

1. Introduction

Meerkats

- Social structure : Live in cooperative groups and work together for everyday tasks.
- Adaptability : Adapted to live in harsh environments.
- **Communication system :** Diversified vocal repertoire.

https://www.bbcearth.com/news/meerkats-theyre-just-like-us

Calls

- Communication among meerkats occurs through various vocalizations, including barks, chirps, trills, and growls.

- Essential in coordinating group activities, warning of potential dangers, and maintaining social cohesion.

- Researchers have identified and classified around **30** vocalization types in meerkats

Call analysis

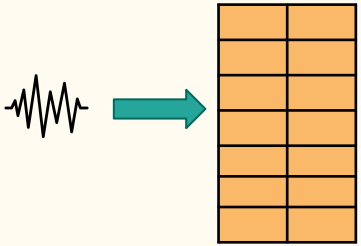
- Improvement in decoding the context of calls → Insights into the social and contextual aspects

Issues ?

- The process of categorizing is conducted by human listeners
- varying in interpretation may arise

Lack of computational methods for the automatic analysis of this language.

In this paper

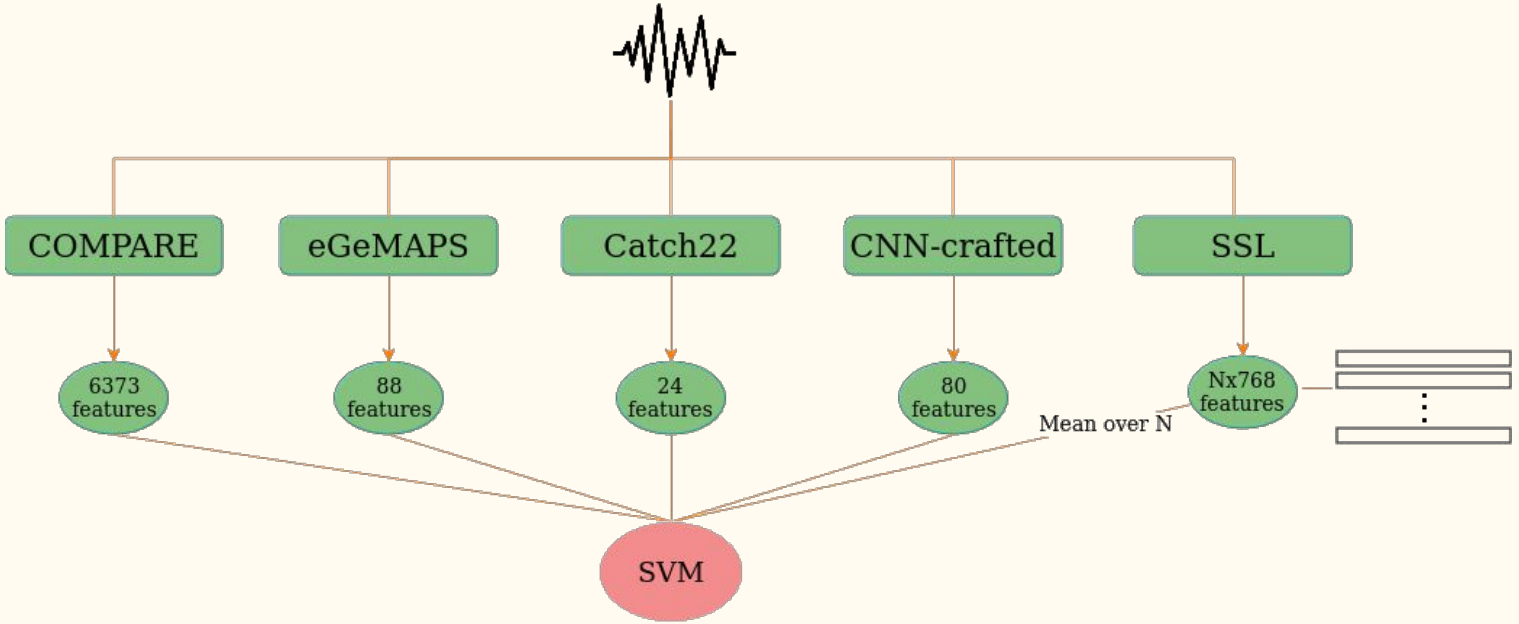


Feature representation plays an important role in pattern analysis and classification systems

- ➢ combining prior knowledge with signal processing
- ➢ data-driven feature representation have become more prominent and useful

❑ Goal : investigate feature representation for automatic meerkat vocalization analysis.

2. Features representation


COMPARE
eGeMAPS
Catch22
CNN-crafted
SSL
6373 features
88 features
24 features
80 features
Nx768 features
Mean over N
SVM

Knowledge-based feature representations : Catch22

- CAnonical Time-series CHaracteristics[1] features are a subset of Highly Comparable Time-Series Analysis :
  - 7700 features extracted by characterizing the signal by different time series analysis methods ⇨ Linear correlation, modeling fitting and etc
- Subset of features that are minimally redundant, tested across 93 real-world time-series classification problems
- Final size of 24 with mean and standard deviation

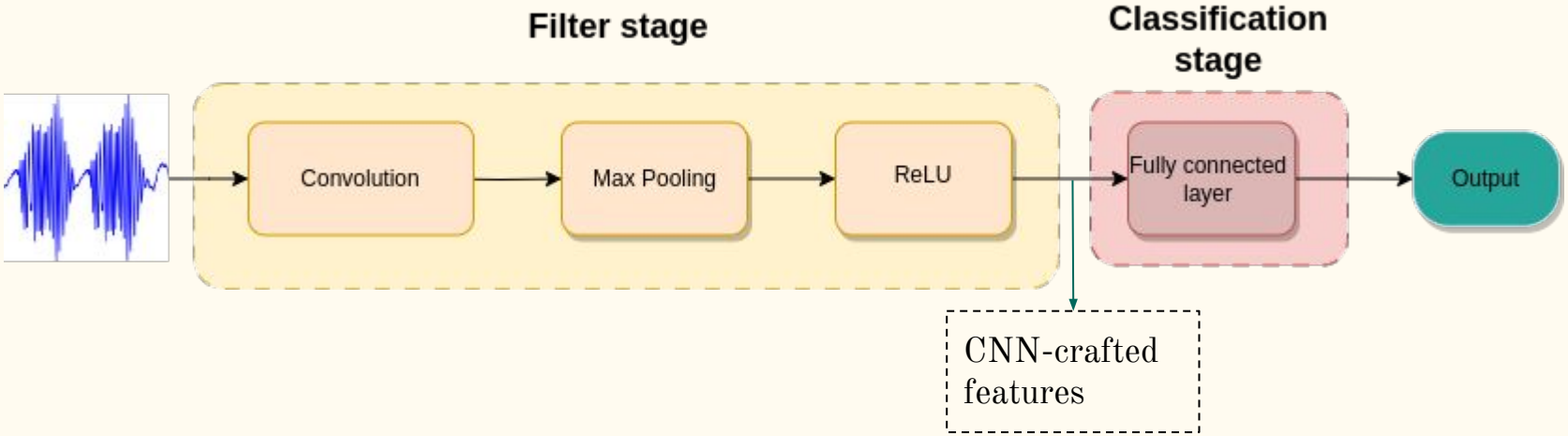
[1]Lubba et al., *catch22: Canonical time-series characteristics*, Data mining and knowledge Discovery, 2019

Knowledge-based feature representations : COMPARE & eGeMAPS

- For paralinguistic speech processing
- COMPARE[1] : 6373 functionals of energy related low level descriptors, spectral LLDs and voicing related LLDS estimated over an utterance
- eGeMAPS : 88 frequency-related parameters, energy/amplitude, spectral and temporal features.

[1]Schuller and al, *The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language*, Interspeech, 2016
[2]Eyben and al, *The geneva minimalistic acoustic parameter set for voice research and affective computing*, IEEE transactions on Affective Computing, 2016

Neural based/data-driven feature representations : supervised learning-based



- Output of the penultimate layer of the model is taken as a feature set, 80 features.

Neural based data-driven feature representations : Self-supervised learning-based

- 3 models used : WavLM[1], wav2vec2[2], HuBERT[3]. With CNN encoder and 12 layers transformer blocks

Model	Corpus
wav2vec2	Librispeech 960
WavLM	Librispeech 960
HuBERT	Librispeech 960

[1]Chen and al, *Wavlm: Large-scale self-supervised pre-training for full stack speech processing*, 2022.
[2]Baevski and al, *wav2vec 2.0: A framework for self-supervised learning of speech representations*, 2020
[3]Hsu and al, *Hubert: Self-supervised speech representation learning by masked prediction of hidden units*, 2021

3. Experiments

Datasets

Set A

- 1795 calls
- 290 s
- 9 categories

Set B[1]

- 6428 calls
- 954 s
- 7 categories

aggression	sentinel	alarm	chatter	grooming	close-call	submission	lead	sunning
125	411	609	108	12	81	99	28	322

aggression	close-call	alarm	lead	short note	social	move
375	1477	645	164	1854	1154	759

[1] Thomas and al, *A practical guide for generating unsupervised spectrogram-based latent space representations of animal vocalizations*, Journal of Animal Ecology, 2022

Experimental set-up

- Preprocessing : Downsample to 16 kHz, minimum 100 ms call length.
- Extraction :

  (a) pycatch22 toolkit was employed call-level Catch22 features

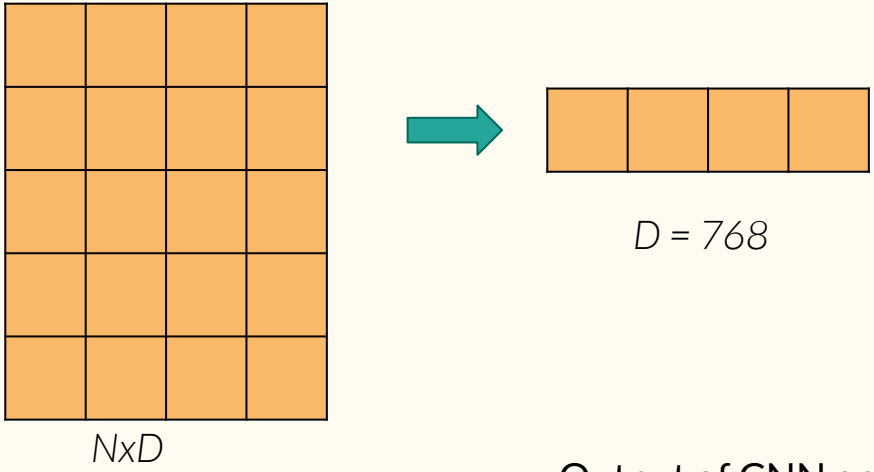
  (b) openSMILE[1] tool is used to extract COMPARE and eGeMAPS feature representations.

  (c) Stratified 5 k-folds strategy to get a CNN feature extractor

  (d) SSL feature representation

[1]Eyben and al, *Opensmile: the munich versatile and fast open-source audio feature extractor*, 2010

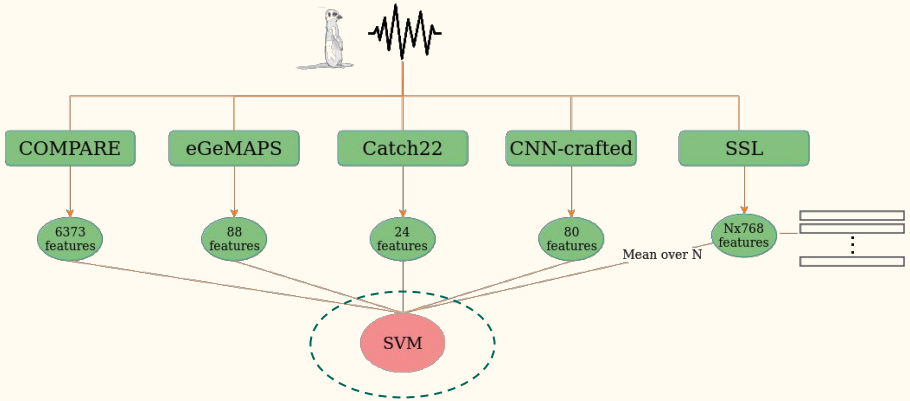
SSL feature representation



- Output of CNN encoder, 1st , 2nd, 6th and last transformer layer.
- Average over the 12 layers per frame and then averaged over frames.
- Use of S3PRL[1] toolkit

[1]Wen Yang and al, SUPERB: Speech Processing Universal PERformance Benchmark, Interspeech, 2021

Classification



- a 5-fold cross-validation grid search strategy by employing 80:20 train-test split

- Use of unweighted average recall as evaluation metric

4. Results and discussion

SSL neural embeddings

	Set A			Set B		
Model	Wav2vec2	WavLM	HuBeRT	Wav2vec2	WavLM	HuBeRT
CNN	0.71	0.68	**0.74**	0.78	0.77	0.78
1st Transformer	0.71	**0.72**	0.73	0.79	0.82	0.78
2ndTransformer	**0.73**	0.71	0.72	**0.79**	**0.82**	**0.79**
6th Transformer	0.54	0.50	0.64	0.69	0.70	0.76
Last Transformer	0.35	0.38	0.55	0.52	0.53	0.67
Average of transformers	0.63	0.59	0.61	0.75	0.72	0.76

- Lower layer transformer and CNN encode yield better systems than higher layers.

Results

- eGeMAPS and COMPARE feature based systems yield better system than Catch22 feature representation.
- In the case of SSL feature representations, the systems are comparable.
- The CNN-crafted feature representation yields the best systems.

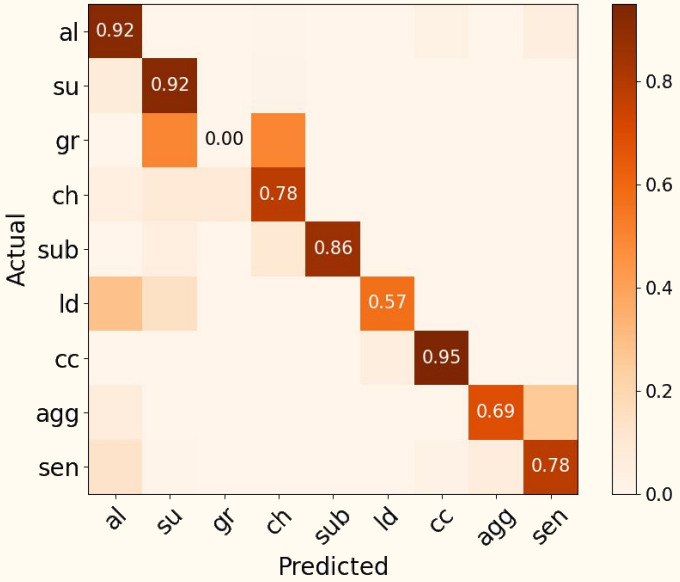
Model	Set A	Set B
eGeMAPS	0.61	0.66
COMPARE	0.80	0.75
Catch22	0.61	0.56
wav2vec2	0.73	0.79
WavLM	0.72	0.82
HuBERT	0.74	0.79
CNN-crafted	0.84	0.84

Discussion

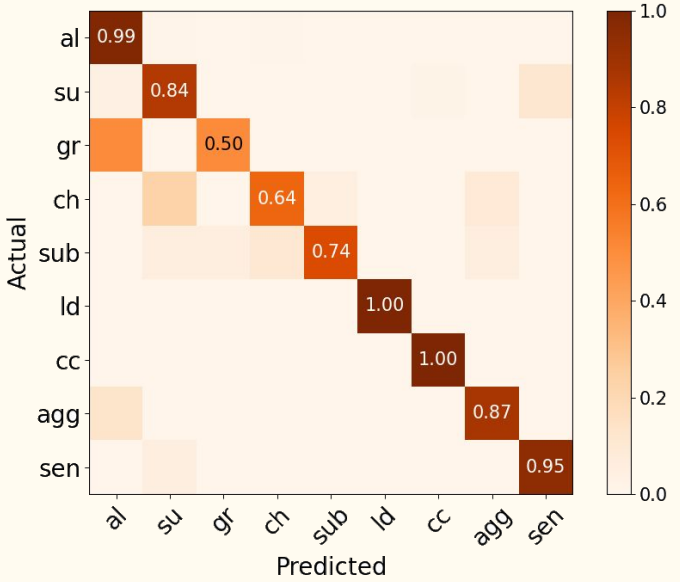
- Similar to neural embeddings from networks pre-trained on human speech, hand-crafted representations developed for speech processing applications can be useful for meerkat call classification.

Confusion matrix : Set A

WavLM

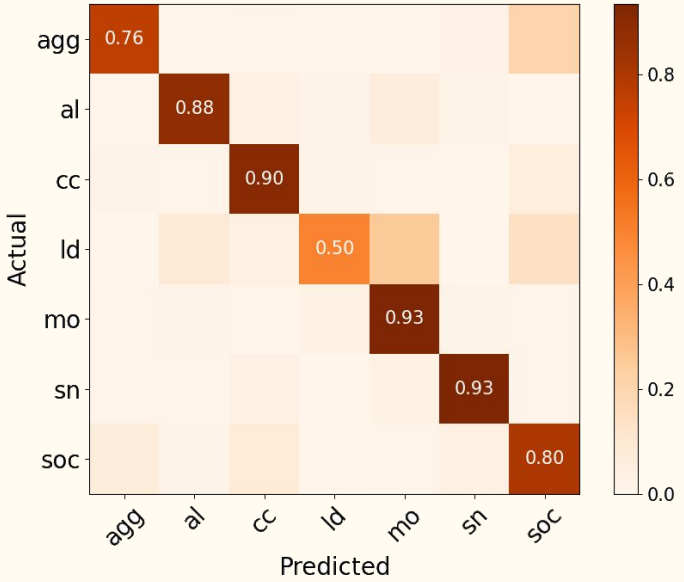


CNN-crafted

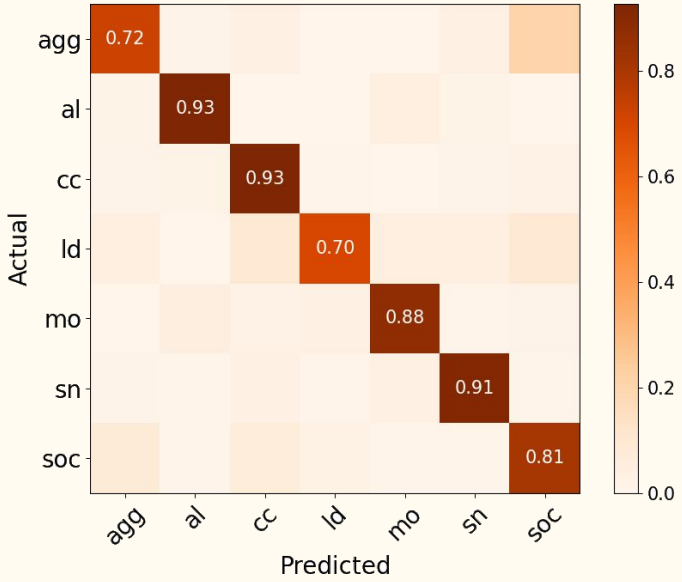


Confusion Matrix : Set B

WavLM



CNN-crafted



5. Conclusion

Conclusion (I)

🐦 Meerkats provide an intriguing model system for investigating animal communication

🐜 Challenge: Lack of methods for automatic meerkat call analysis.

➢ In that direction, this paper explored feature representations for automatic analysis of meerkat vocalizations.

🐸 We compared time series analysis-based hand-crafted feature representation, hand-crafted feature representations, SSL-based feature representations obtained from neural networks, and feature representations automatically learned in a task-dependent manner from meerkat calls using CNNs.

Conclusion (II)

🐙 Our studies show that hand-crafted feature extractors and SSL feature extractors developed for human speech processing can be used for meerkat call classification.

🐒 CNN-based method developed for automatic feature learning in a task-dependent manner for human speech processing can be scaled for meerkat call classification task (CNN-crafted).

🐋 Our future work will focus on analyzing these diverse feature representations to tease out and explain the acoustic information that is relevant for meerkat call analysis.

Thank you!

Neural based data-driven feature representations : Self-supervised learning-based

- Need of labeled data for traditional supervised learning ⇒ Expensive and time-consuming

Solution ❓

- Self Supervised Learning : Leverage these unlabeled data and design pretext task as training criterion.
- meaningful representation are learned and can be used for a downstream task using labeled data
- 3 models used : WavLM, wav2vec2, HuBERT. With CNN encoder and 12 layers transformer blocks