Towards Leveraging Sequential Structure in Animal Vocalizations

Eklavya Sarkar[1,2], Mathew Magimai Doss[1]

[1]Idiap Research Institute [2]École Polytechnique Fédérale de Lausanne

Sequential Structure in Animal Vocalizations

- Animal calls are complex and structured sequences of acoustic units.
- Often combined in species-specific ways, following syntactic rules.

Common Problem in Bioacoustics Classification

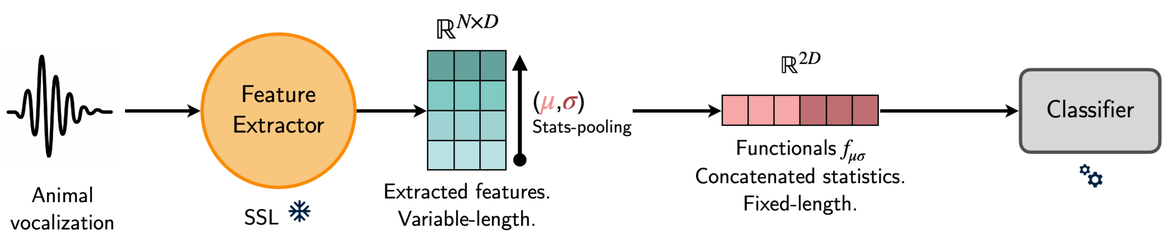
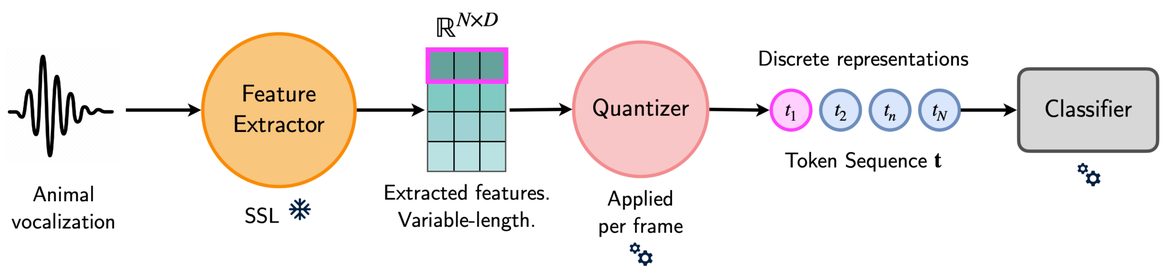


Figure 1. Each vocalization treated like an unordered collection of frame-level features.



- **Goal**: investigate alternate feature representations that can capture the sequential structure within animal vocalizations, and leverage the unutilized temporal information to improve classification performance.

Discrete Audio Tokens

Quantizer	Hyperparameter	Search Space
VQ	Learning rate	1e[-4, -3, -2]
	Commitment cost	0.25
	EMA	[True, False]
GVQ	Learning rate	1e[-4, -3, -2]
	KL weight	[0.5, 1.0, 1.5, 2.0]
	Diversity weight	[0.0, 0.01, 0.05, 0.1, 0.2, 0.5]
	Temperature schedule:	
	Max temperature	2.0
	Min temperature	0.1
	Decay factor	0.999

Table 1. Hyperparameter search space for VQ and GVQ models.

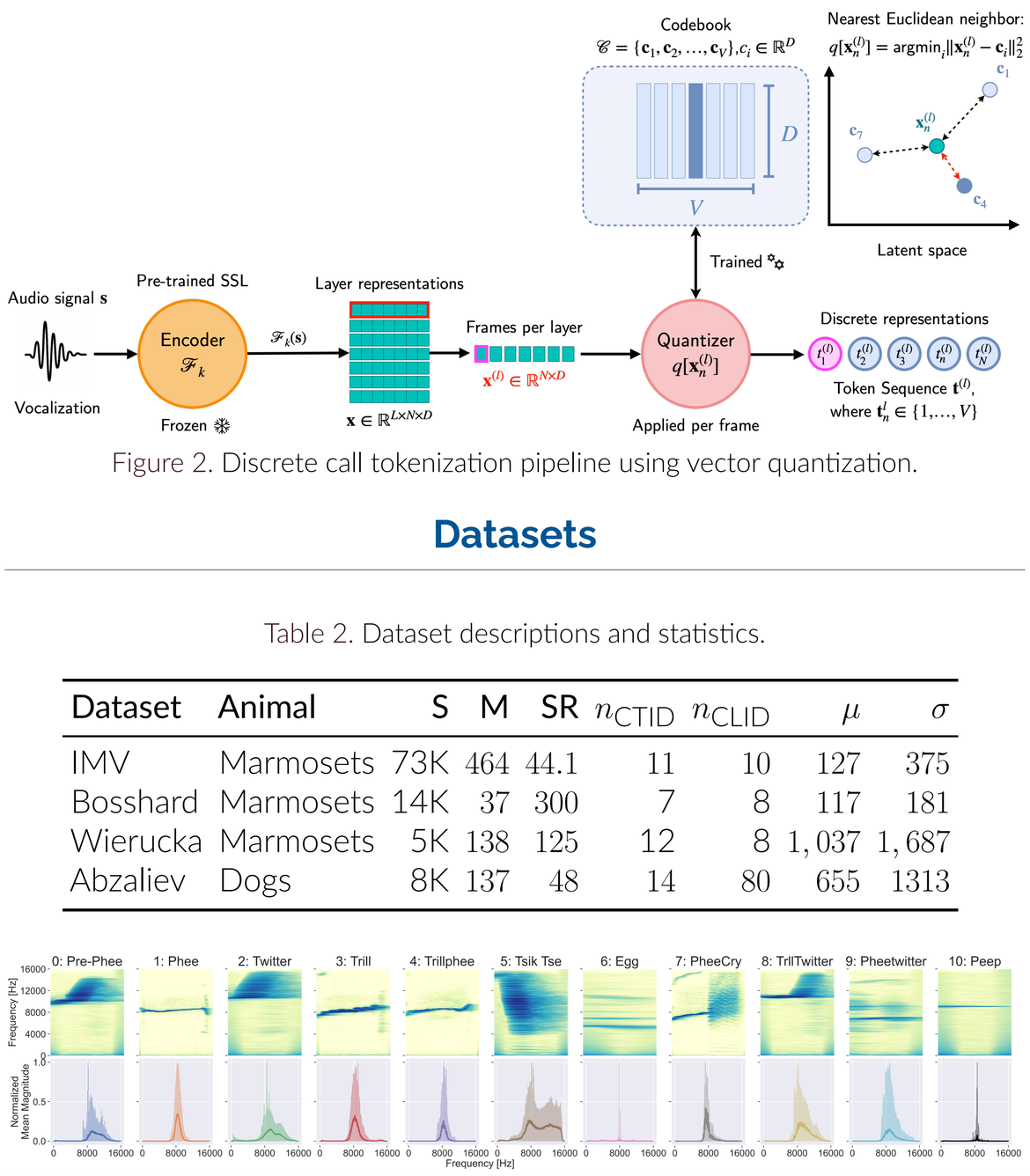


Figure 2. Discrete call tokenization pipeline using vector quantization.

Datasets

Table 2. Dataset descriptions and statistics.

Dataset	Animal	S	M	SR	n_{CTID}	n_{CLID}	μ	σ
IMV	Marmosets	73K	464	44.1	11	10	127	375
Bosshard	Marmosets	14K	37	300	7	8	117	181
Wierucka	Marmosets	5K	138	125	12	8	1,037	1,687
Abzaliev	Dogs	8K	137	48	14	80	655	1313

Levenshtein Distance Analysis

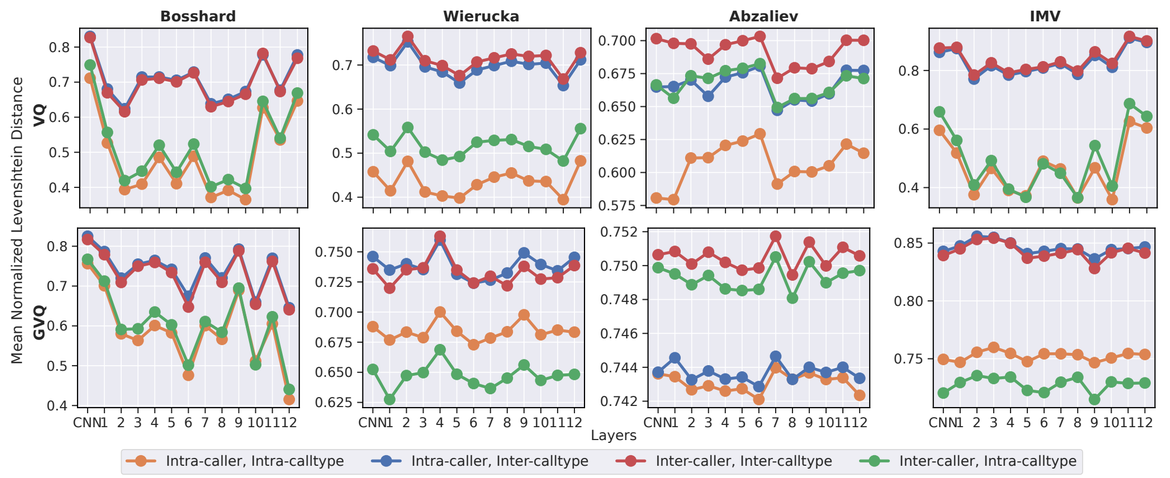


Figure 4. Layer-wise mean Levenshtein distance between all pairs of token sequences.

- VQ tokens are indeed capable of clustering sufficient acoustic information to discriminate calls or callers, under real-world left-to-right temporal constraints.
- The GVQ tokens exhibit some unexpected patterns and less consistent separability, indicating they may be less effective.

k-NN Classification Analysis

Table 3. Hyperparameter search space used for training the k-NN classifier.

Classifier	Hyperparameter	Search Space
k-NN	Number of neighbours k	[1, 3, 5, 7, 9]
	Neighbour weighting	[Uniform, distance]
	Distance	Levenshtein
	Task	[CTID, CLID]

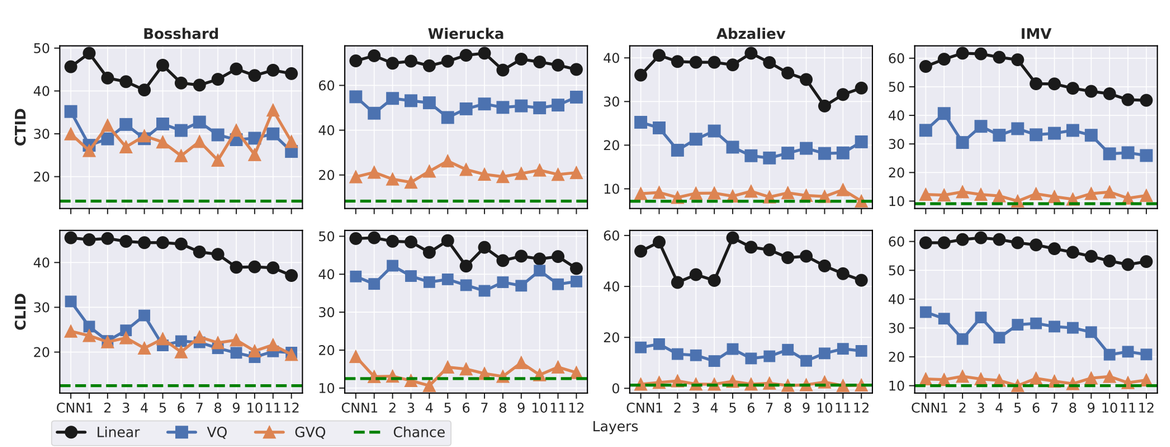


Figure 5. Layer-wise UAR [%] using k-NN on token sequences.

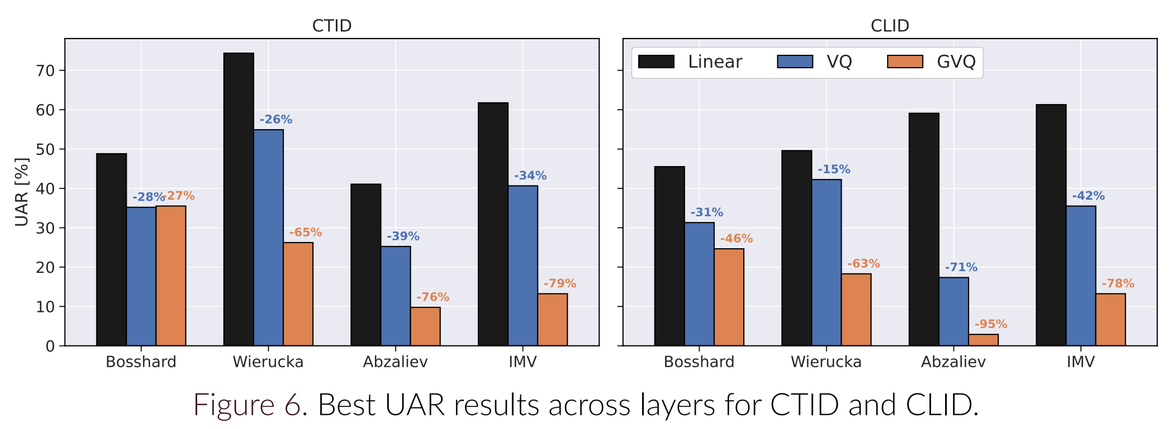


Figure 6. Best UAR results across layers for CTID and CLID.

Conclusions

- Discrete acoustic **token** sequences are able to **encode temporal information**, and exhibit a **degree of separability** by call-type or caller identity across all datasets.
- k-NN classifier showed that while VQ token representations are still weaker than linear-probing baselines, they are nonetheless able to **leverage meaningful sequential information** from animal vocalizations.
- Incorporating **more sophisticated** sequence **modeling** could further improve performance.

Acknowledgement

This work was funded by the Swiss National Science Foundation's *NCCR Evolving Language* through grant agreement no. 51NF40 180888.