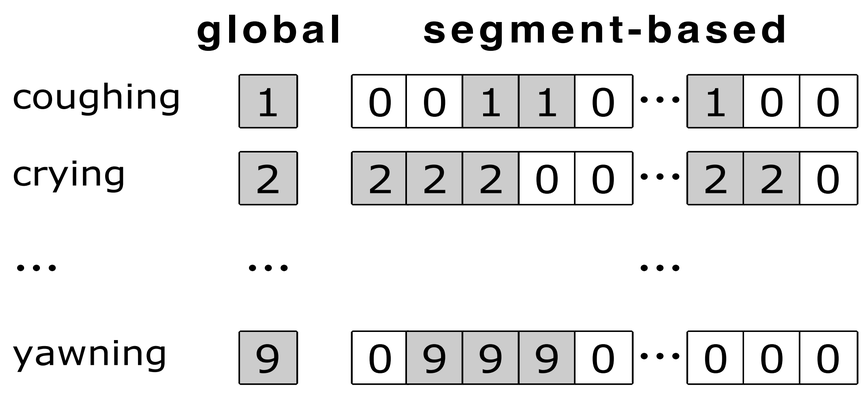
Controlling Multi-Class Human Vocalization Generation via a Simple Segment-based Labeling Scheme

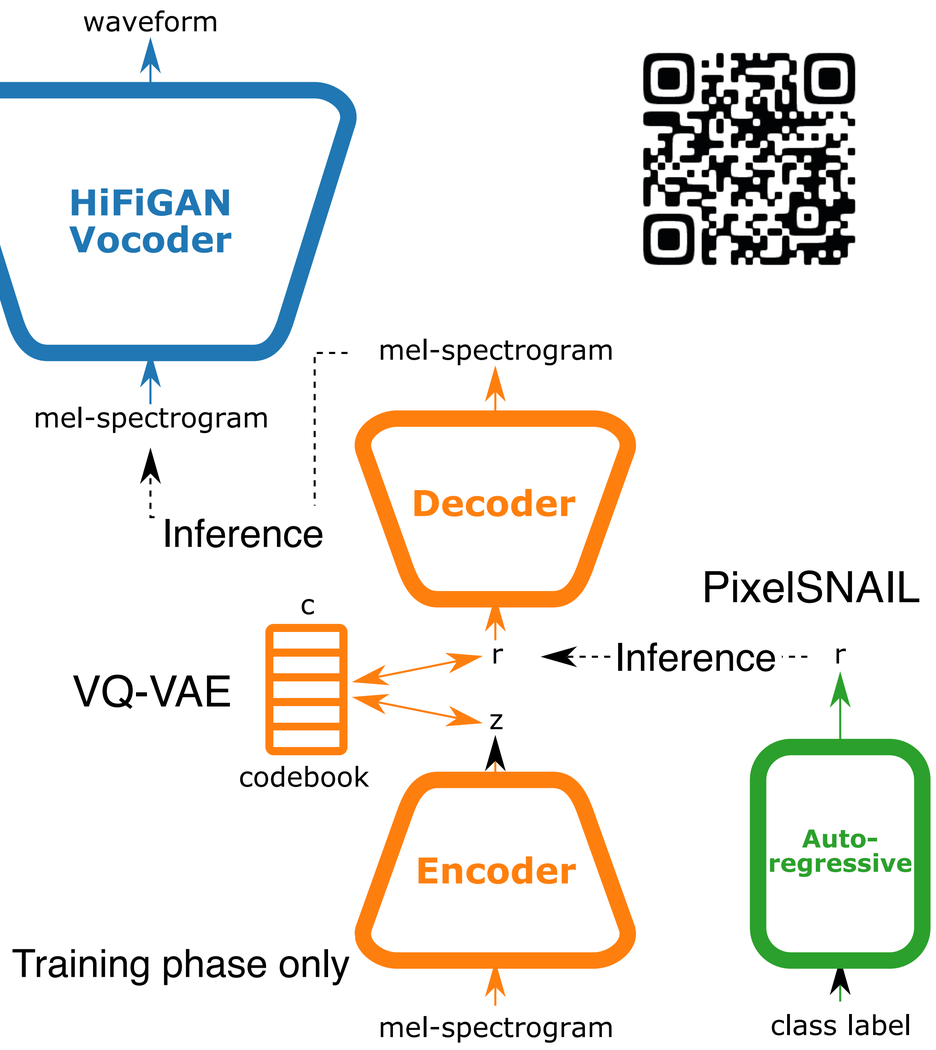
Hieu-Thi Luong & Junichi Yamagishi, National Institute of Informatics, Japan

Abstract

- What is the appropriate unit to model, generate, and control vocalization such as laughing and crying?
- The investigated unit in this paper: Uniform segment units (see the right-side figure)
- Switching the unit from global to segment does not degrade the quality, but does degrade the perception of vocalization classes
- Bringing the controllability of vocalization



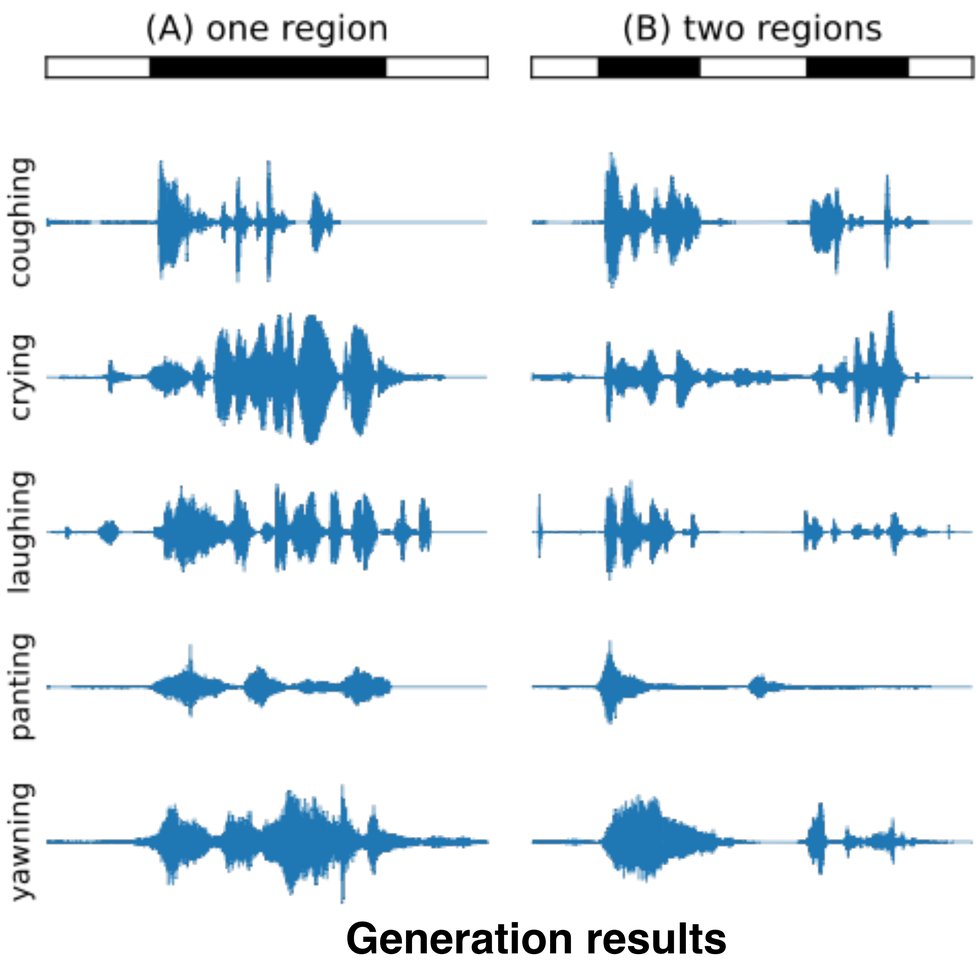
Human vocalization generation system used for this study



Datasets

- *Pretraining (speech):*
  - VCTK
- *Fine-tuning (vocalization)*
  - Deeply Nonverbal Vocalization dataset
    1. coughing
    2. crying
    3. laughing
    4. moaning
    5. panting
    6. screaming
    7. sighing
    8. throat-clearing
    9. yawning

How to annotate segment labels: if the RMS value within its window is above -24 dB

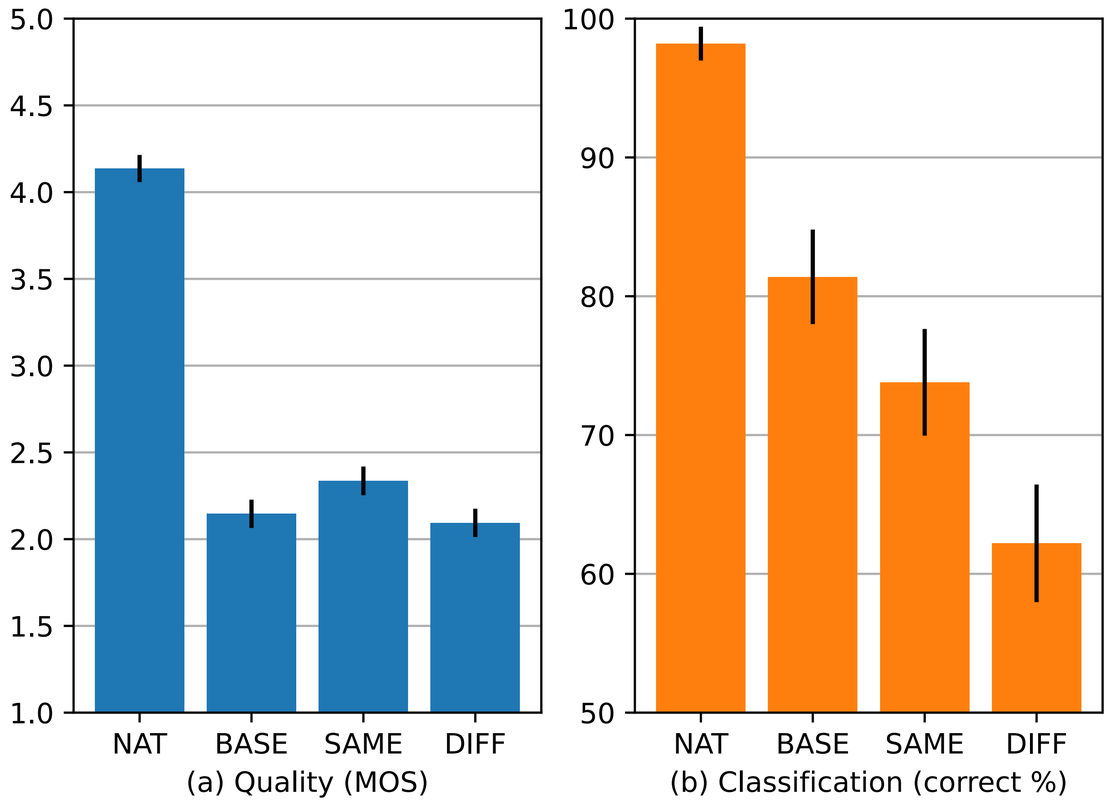


Generation results

Experiments

Experimental systems

- NAT
  - Natural audio
- **BASE**
  - Uses global labels
- **SAME**
  - Uses segment labels extracted from natural samples of the same vocal class
  - e.g. segment-based labels extracted from natural laughing samples to generate laughing samples
- **DIFF**
  - Uses segment labels extracted from natural samples of the different vocal class
  - e.g. segment-based labels extracted from coughing, crying, panting, yawning, and screaming to generate laughing samples



Listening test results (10 subjects)

(a) BASE

Class	MOS	Classification (%)				
		1	2	3	4	5
1. Coughing	2.44	**80.0**	12.0	7.0	0.0	1.0
2. Crying	2.01	2.0	**82.0**	12.0	2.0	2.0
3. Laughing	2.12	1.0	11.0	**86.0**	0.0	2.0
4. Panting	1.88	6.0	9.0	0.0	**71.0**	14.0
5. Yawning	2.28	5.0	1.0	0.0	6.0	**88.0**

(b) SAME

Class	MOS	Classification (%)				
		1	2	3	4	5
1. Coughing	2.73	**92.0**	3.0	3.0	1.0	1.0
2. Crying	2.21	0.0	**64.0**	16.0	3.0	17.0
3. Laughing	2.13	2.0	15.0	**81.0**	2.0	0.0
4. Panting	2.20	3.0	18.0	3.0	**52.0**	24.0
5. Yawning	2.41	6.0	2.0	1.0	11.0	**80.0**

(c) DIFF

Class	MOS	Classification (%)				
		1	2	3	4	5
1. Coughing	2.31	**75.0**	3.0	21.0	0.0	1.0
2. Crying	2.17	5.0	**61.0**	17.0	3.0	14.0
3. Laughing	2.14	3.0	29.0	**63.0**	2.0	3.0
4. Panting	1.74	7.0	20.0	4.0	**31.0**	38.0
5. Yawning	2.11	2.0	9.0	0.0	8.0	**81.0**

Confusion matrix of the classification results

Contact: jyamagis@nii.ac.jp