Quantizer robustness training in Generative Spoken Language Modeling

Raphaël Bernas
ENSTA-MVA
raphael.bernas@ensta.fr

Maxime Corlay
ENSTA-MVA
maxime.corlay@ensta.fr

Adrien Letellier
ENSAE-MVA
adrien.letellier@ensae.fr

Emilie Zheng
ENS-MVA
emilie.zheng@ens.psl.eu

I. Introduction : Robustness of Speech-to-Unit Models

- Invariance of the units from the speech-to-unit block to non-spoken augmentations

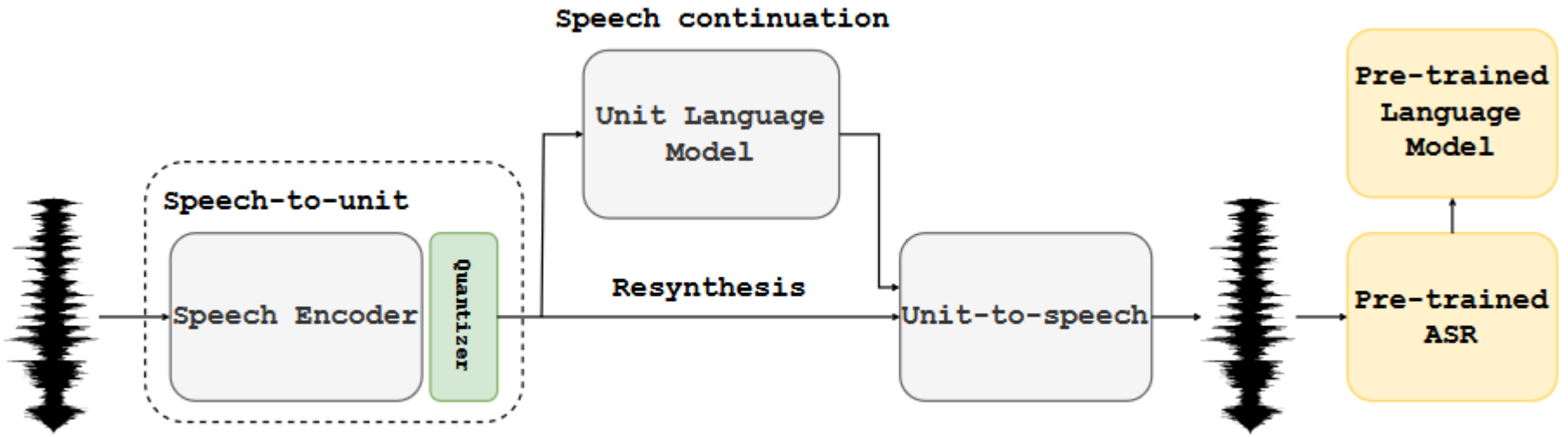


Figure 1: Generative Spoken Language Modeling is composed of three components: (i) Speech-to-unit, (ii) Unit language model, and (iii) Unit-to-speech. Pre-trained ASR and language models are used for evaluation.

II. Metrics to evaluate the Robustness

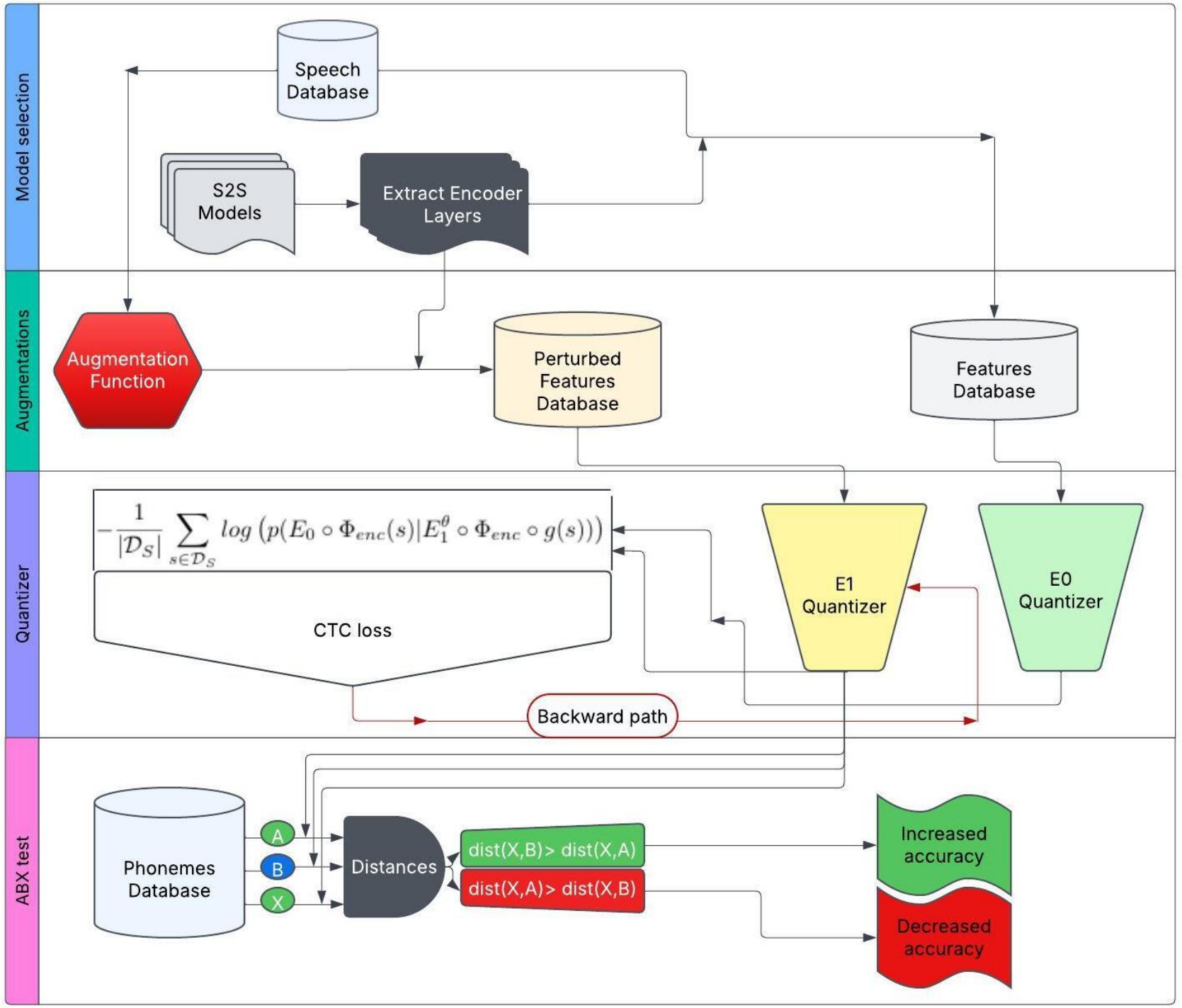
- **Unit Edit Distance**: a Levenshtein distance applied to the deduplicated discrete representations normalized by the number of frames

$$\sum_{x \in \mathcal{D}} \frac{1}{T'_x} \mathrm{LEV}\left((E \circ f)(x), (E \circ f \circ g)(x)\right)$$

-> A robust model should return similar outputs for non-augmented and augmented inputs

- Alternative metrics that takes the embedding space into account: **Dynamic Time Warping**

III. General pipeline



Source : Raphael Bernas using LucidChart

IV.
Augmentation

- Gaussian noise (as in the original paper)
- Time stretch (as in the original paper)
- Pitch Shift (as in the original paper)
- Clipping
- Lowpass, Bandpass, Highpass
- Little bips, Big bips, Corruption
- Echo (as in the original paper)
- Changing volume

V. Quantizer outputs for UED computation

	UED E0	UED E1	UED E2
Missing samples	1.57	2.29	2.23
Volume change	0.42	0.40	0.29
Noise adding	4.02	5.26	5.12
Pitch shifting	1.85	3.67	2.70

ABX Evaluation	
E1	E2
0.39	0.38

VI. Testing new models

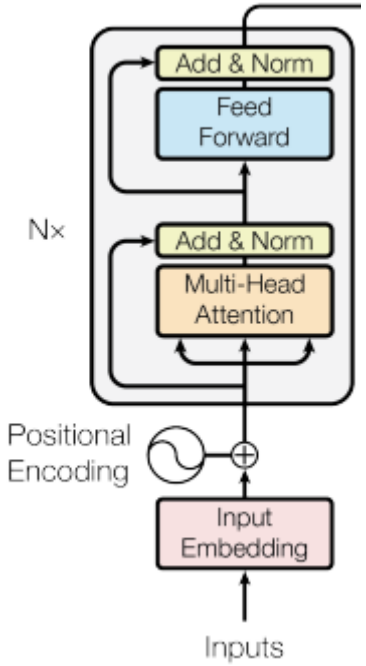
	UED E0	UED E1	UED E2
Noise adding	1.72	2.59	4.35
Time stretch	2.48	3.10	4.55
Pitch shifting	2.36	3.34	2.1
Clipping	5.28	6.06	4.05

ABX Evaluation	
E1	E2
0.32	0.33

Table 3. UED scores for the quantizers E0 (kmeans), E1 and E2, and ABX scores for E1 and E2 trained on HuBERT Large.

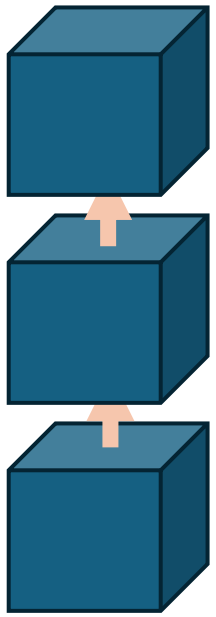
VII. Testing new Quantizer architecture

Transformers



VS

Big MLP



Source: « Attention is all you need » Vaswani et al.

Augmentation	E0	E1B	E1T
gaussian_noise	3.03896	3.36652	5.24032
time_stretch	2.00196	4.45919	4.60026
pitch_shift	1.84524	3.98918	4.23216
clipping	6.24134	3.70988	6.28224
lowpass	6.07134	4.04882	5.63811
bandpass	7.37028	4.23944	5.90323
highpass	2.22209	3.02597	4.72911
little_bips	5.21451	3.99738	6.18291
big_bips	0.76819	2.41804	2.75212
echo	3.10572	3.33085	4.97011
corruption	0.92529	2.39520	2.72410
volume_change	0.00000	2.21813	0.00000

ABX Evaluation	
E1B	E1T
0.37	0.33

Conclusion

- Variable effectiveness depending on the types of audio distortions
- Strong dependency of results on the specific run
- Significant impact of encoder model size on results across different augmentations
- Potentially limited reliability of results due to small dataset size (50 samples)