Textless NLP

吳冠緯、詹侑昕

Outlines

- Introduction to Textless NLP
- Paper survey
- SQA
- Future works
- References

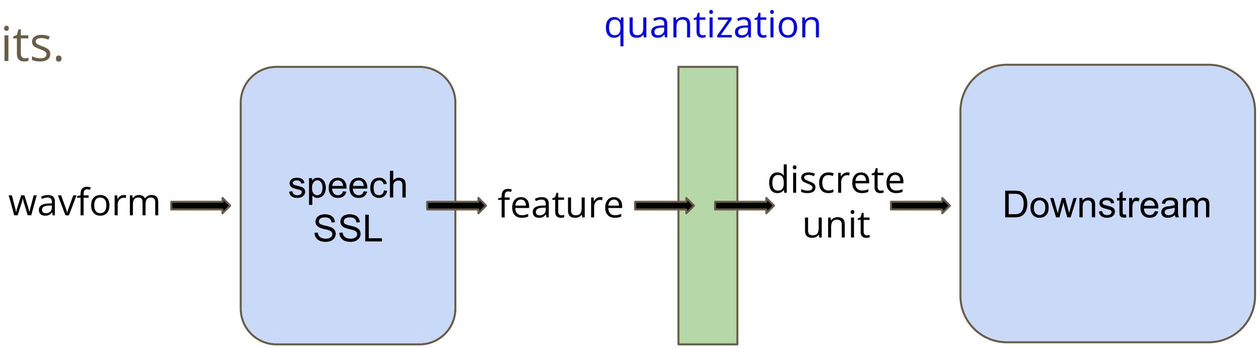
Introduction to Textless NLP

Textless NLP

- Applying language model directly to audio inputs, side stepping the need for textual resources or ASR. (Escaping from the potential error of ASR.)
- Beneficial for languages which do not have large textual resources or a widely used standardized orthography.
- Some linguistically relevant signals carried by prosody and intonation are basically absent from text.

Speech to Discrete Unit

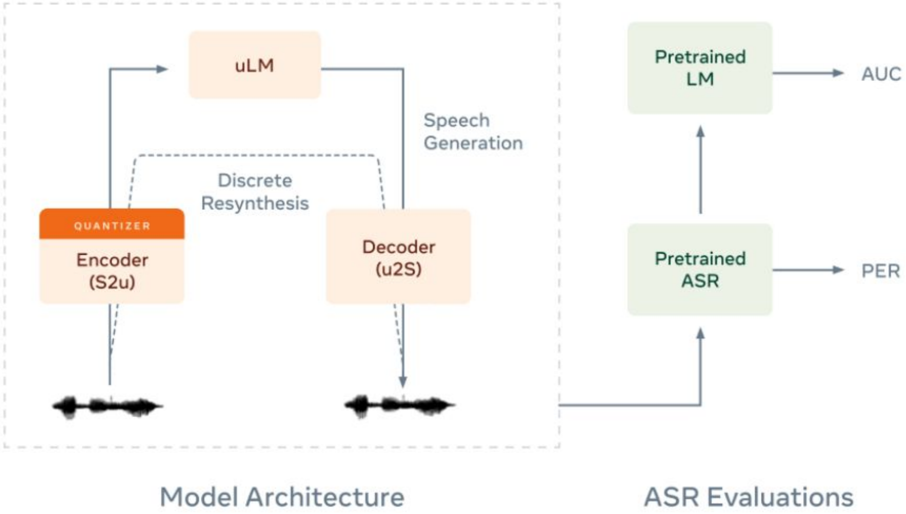
- We can find out informative speech representations due to the success of self-supervised speech pre-training.
- We apply quantization on those speech representations and discover discrete speech units.



Paper survey

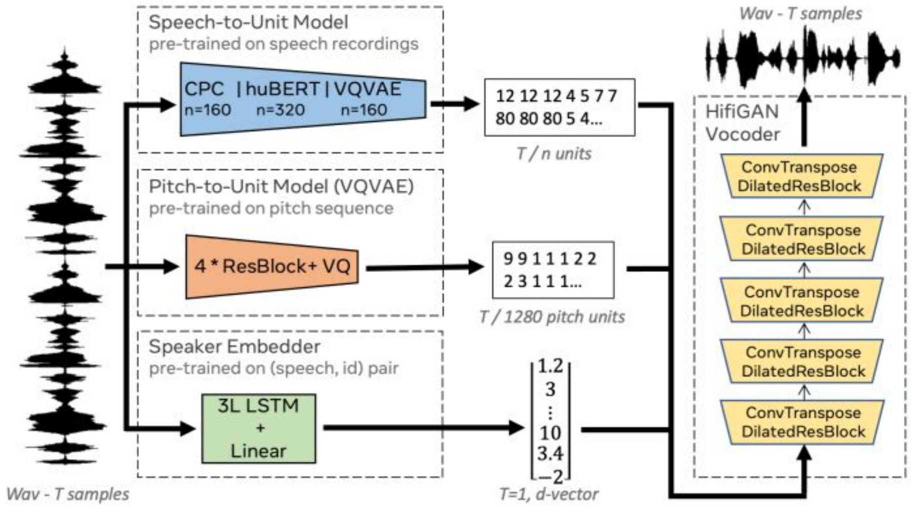
Generative Spoken Language Model

- GSLM begins by building a baseline model and evaluating it on two simple end-to-end tasks.
- The language model were trained on the discrete units (pseudo-text) from raw audio.



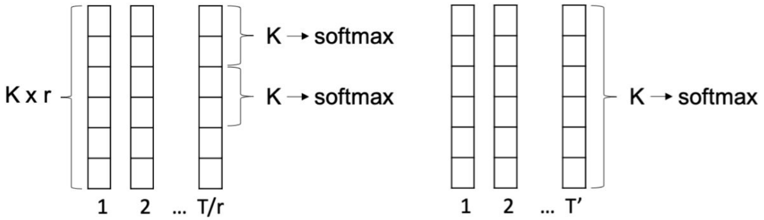
Speech Resynthesis from Discrete

- Using discrete units as the disentangled representations for speech resynthesis.
- Capturing prosody by improving the encoder and decoder.



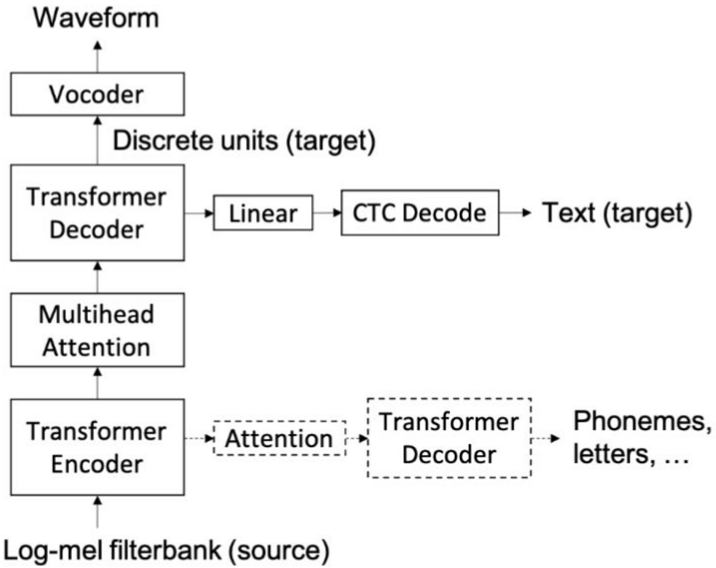
Speech to speech translation with discrete units

- Using a Speech-to-unit model to generate discrete units



(a) *stacked* (b) *reduced*

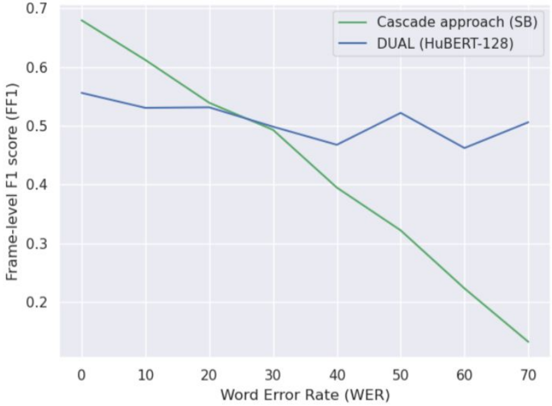
- Incorporating auxiliary tasks with additional attention and decoder modules



SQA

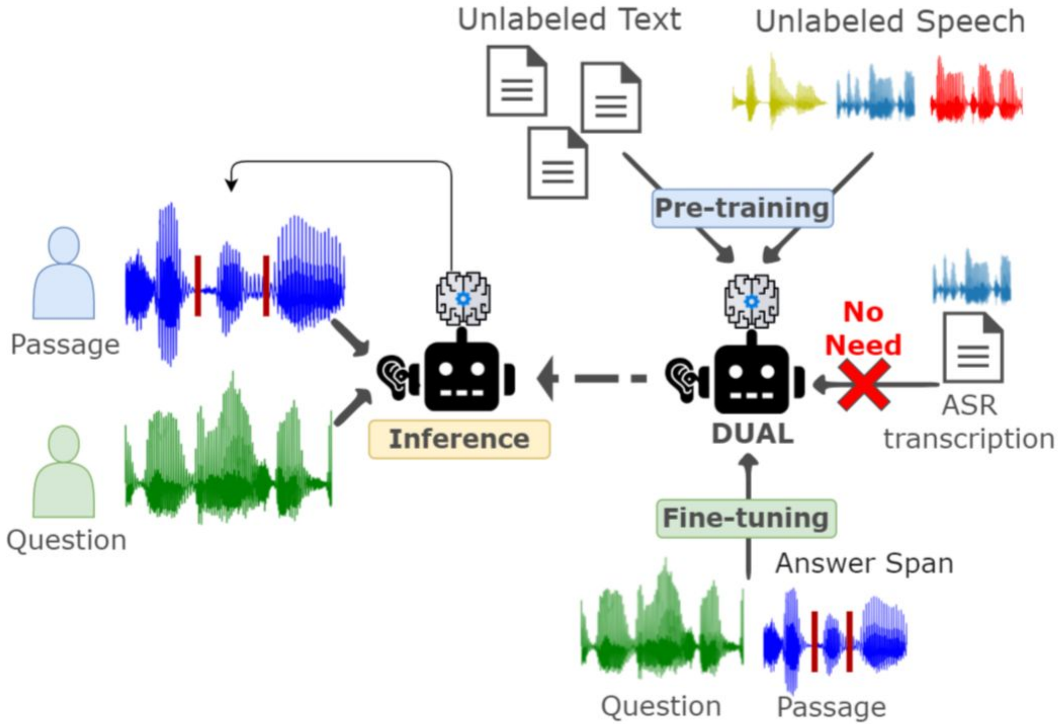
Introduction

- Previous Spoken Question Answering (SQA) works rely on **ASR transcriptions.**
- Drawbacks:
  - Doesn’t work for languages without text
  - ASR errors may lead to catastrophic results
- Task:
  - find the answer span in passage
- Dataset: NMSQA

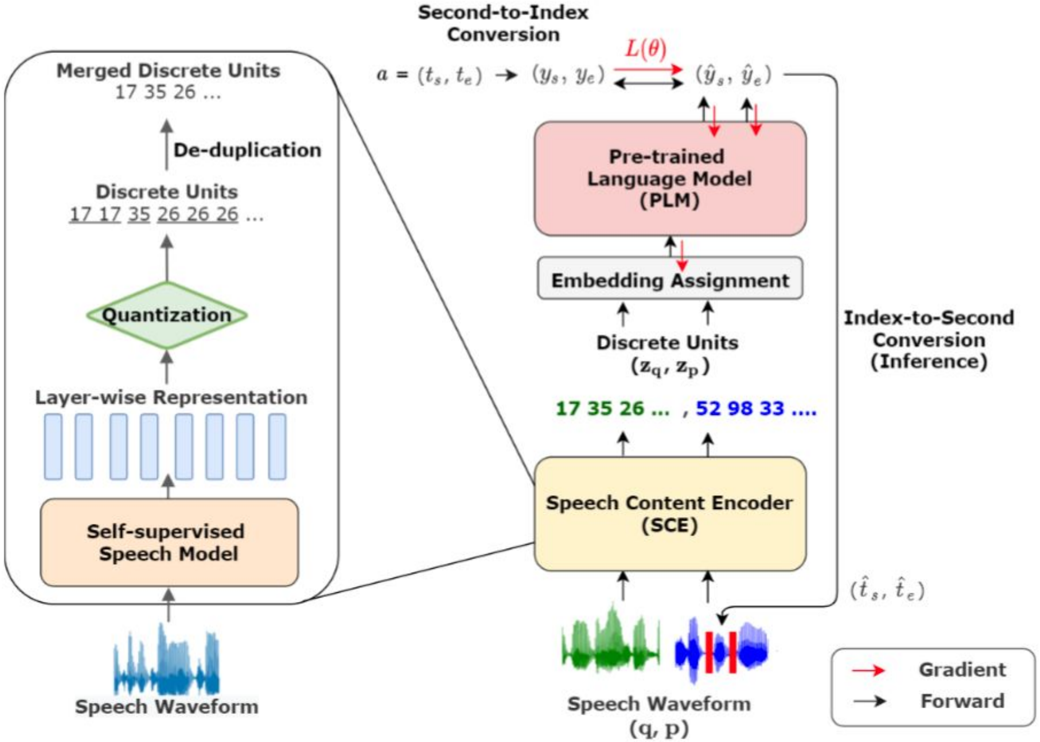


Property	**train**	**dev**	**test-SQuAD**	**test-OOD**
# of Sample	95024	21199	101	166
Hour	297.18	37.61	2.61	8.36
# of Speaker	12	12	60	60
Real Speaker	✕	✕	✔	✔
Content Source	SQuAD-train	SQuAD-dev-1	SQuAD-dev-2	NewsQA-dev, QuAC-dev
Speech Quality	Natural, Clean	Natural, Clean	Disfluent, Noisy	Disfluent, Noisy

Workflow


Unlabeled Text
Unlabeled Speech
Pre-training
Passage
Question
Inference
No
Need
ASR
transcription
DUAL
Fine-tuning
Answer Span
Question
Passage

DUAL Framework



Discrete Unit

- Self-supervised Speech Model
  - Use HuBERT to encode raw waveform into frame-level 1024 dimension features
  - 20ms/Frame
- Speech Quantization
  - Use K-means clustering to cluster features into discrete units
  - deduplicate

Discrete Units: 17 17 35 26 26 26 …

17 35 26…

Pre-trained Language Model

- Input:
  - concatenated discrete units of question and passage pair($\mathbf{z}_q$, $\mathbf{z}_p$)
- Target:
  - start and end position (y_s, y_e) after the deduplication process
- Model:
  - Longformer

Results

Input	Model	dev		test-SQuAD		test-OOD	
		FF1	AOS	FF1	AOS	FF1	AOS
With ASR transcriptions (Cascade Approach)							
ASR prediction (SB)	Longformer†	56.74	49.72	17.34	15.27	16.92	15.66
ASR prediction (W2v2-st-ft)	Longformer†	65.67	58.34	64.17	57.44	57.67	50.31
Without ASR transcriptions (DUAL)							
HuBERT-64	Longformer	47.76	42.22	39.03	32.97	32.58	28.39
HuBERT-128	Longformer	54.22	48.52	**55.93**	**49.13**	**38.63**	**34.61**
HuBERT-512	Longformer	**55.02**	**49.59**	17.28	12.46	10.35	7.40

ASR	LibriSpeech test-clean	NMSQA dev	NMSQA test
SB	3.08	15.75	61.70
W2v2-st-ft	1.90	10.48	11.28

Future works

Test the performance of DUAL in different tasks

- DUAL has shown its ability on SQA
- How about other SLU tasks?
- Our future works
  - NER
  - Intent classification

NER

NER in SLUE Benchmark

Corpus	Size - utts (hour)		
	Fine-tune	Dev	Test
SLUE-VoxPopuli	5000 (14.5)	1753 (5.0)	1842 (4.9)

Speech model	LM	Text model	F1 (%)	label-F1 (%)
NLP Toplines:				
N/A (GT Text)	N/A	DeBERTa-L	81.4	85.7
Pipeline approaches:				
W2V2-B-LS960	-	DeBERTa-L	49.5	74.2
W2V2-L-LL60K	-	DeBERTa-L	57.8	78.8
W2V2-B-LS960	✓	DeBERTa-L	68.0	79.8
W2V2-L-LL60K	✓	DeBERTa-L	69.6	82.2
E2E approaches:				
W2V2-B-LS960	-	N/A	50.2	64.0
W2V2-B-VP100K	-		47.9	60.8
HuBERT-B-LS960	-		49.8	62.9
W2V2-L-LL60K	-		50.9	64.7
W2V2-B-LS960	✓		63.4	71.7
W2V2-B-VP100K	✓		61.8	69.8
HuBERT-B-LS960	✓		61.9	70.3
W2V2-L-LL60K	✓		64.8	73.3

Table 5. Named entity recognition performance on test set.

Intent Classification

Datasets: Smartlights, ATIS

Current Status: data preprocessing

References

References

GSLM: https://arxiv.org/pdf/2102.01192.pdf

speech resynthesis: https://arxiv.org/pdf/2104.00355.pdf

S2S translation: https://arxiv.org/pdf/2107.05604.pdf

SLUE: https://arxiv.org/abs/2111.10367

ATIS: https://aclanthology.org/H90-1021.pdf

Smarlights: https://arxiv.org/pdf/1810.12735.pdf