Spoken Language Modelling without Text

Eugene Kharitonov

facebook AI Research

Spoken Language Modelling without Text

On Generative Spoken Language Modeling from Raw Audio

Kushal Lakhotia*, Eugene Kharitonov*, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte§, Tu-Anh Nguyen†, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, Emmanuel Dupoux‡
Facebook AI Research

Text-Free Prosody-Aware Generative Spoken Language Modeling

Eugene Kharitonov*, Ann Lee*, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu-Anh Nguyen, Morgane Rivière, Abdelrahman Mohamed, Emmanuel Dupoux, Wei-Ning Hsu
Facebook AI Research

Demo page: https://speechbot.github.io/

Practical motivation

Modern NLP techniques operate on textual representations

- However, many languages and dialects have few or no textual resources
- Spoken language is different to written

Current NLP technology leaves them out

- We want to enable NLP to work directly on speech, without texts or labels

But also

- Test-field for language acquisition
- Better understanding language

Examples

- Transformer-based language model, trained on audio w/o text or labels
- We prompt with 3s audio from the test set
- The model auto-regressively generates the continuation

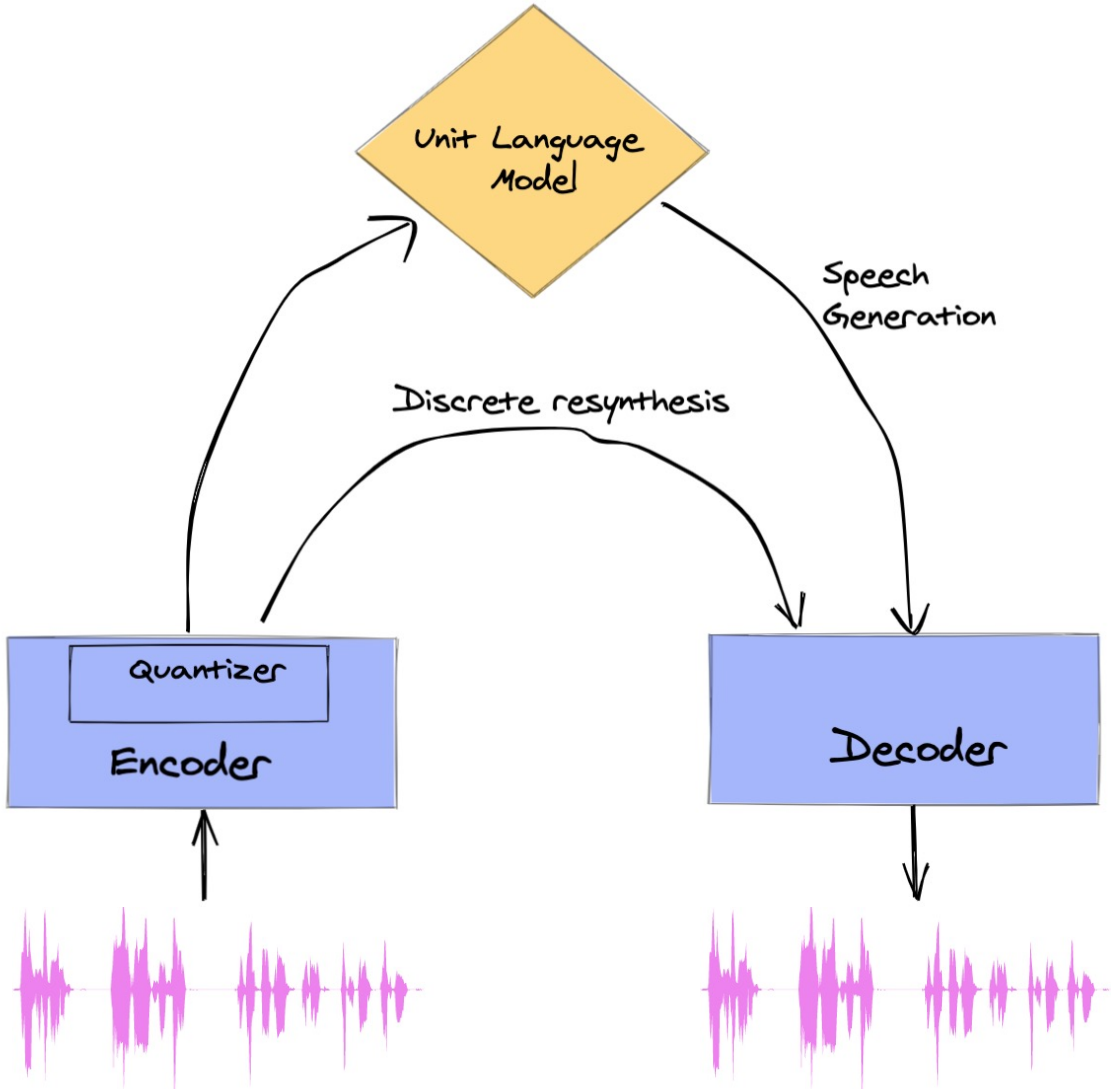
speechbot.github.io/gslm/

Architecture

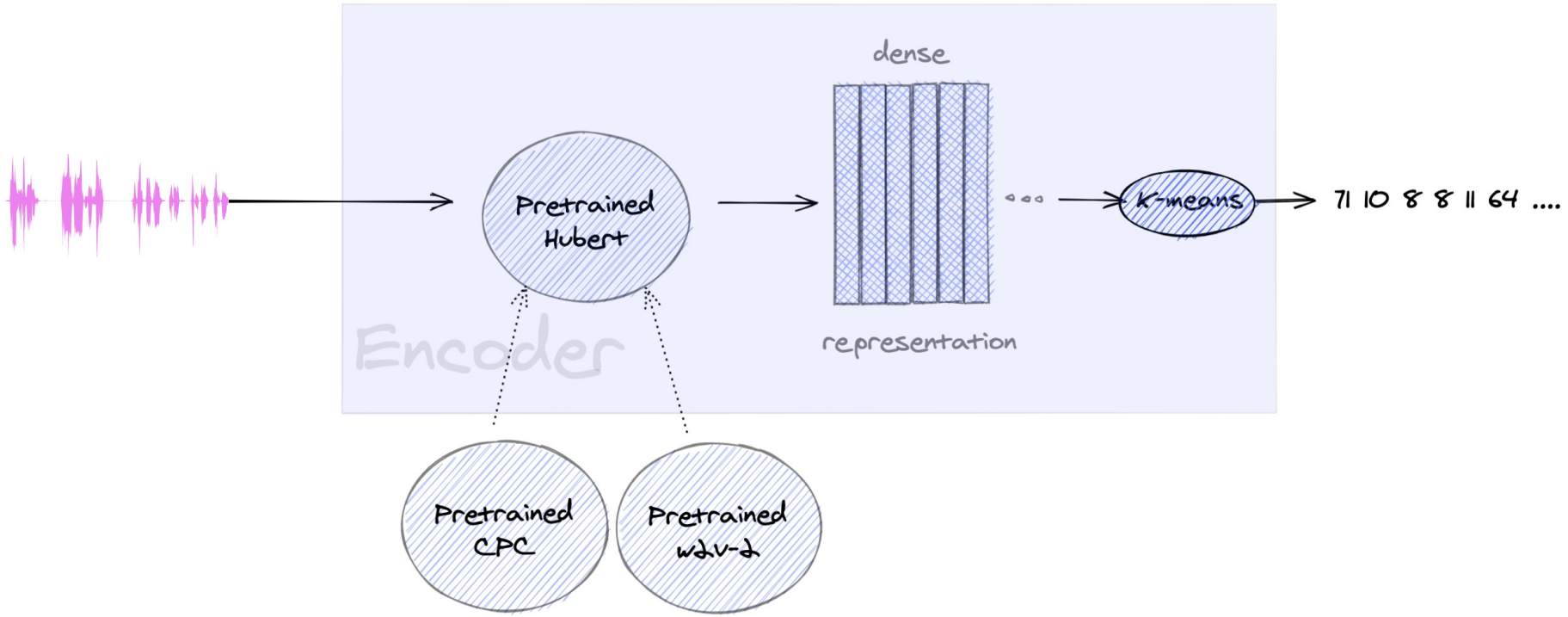
On Generative Spoken Language Modeling from Raw Audio

Kushal Lakhotia*, Eugene Kharitonov*, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte§, Tu-Anh Nguyen†, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, Emmanuel Dupoux‡
Facebook AI Research

Language modelling pipeline overview

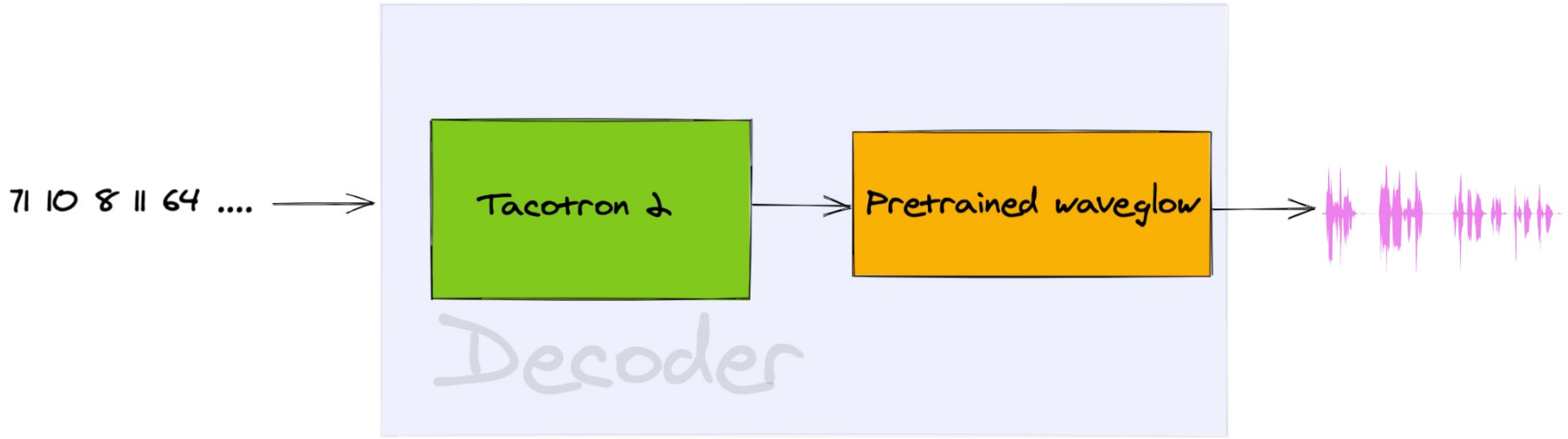


Speech to units: Encoder



HuBERT: How much can a bad teacher benefit ASR pre-training? Hsu et al, 2020
wav2vec 2.0: A framework for self-supervised learning of speech representations, Baevski et al., 2020
Representation learning with contrastive predictive coding, van den Oord, 2018

Units to speech: Decoder



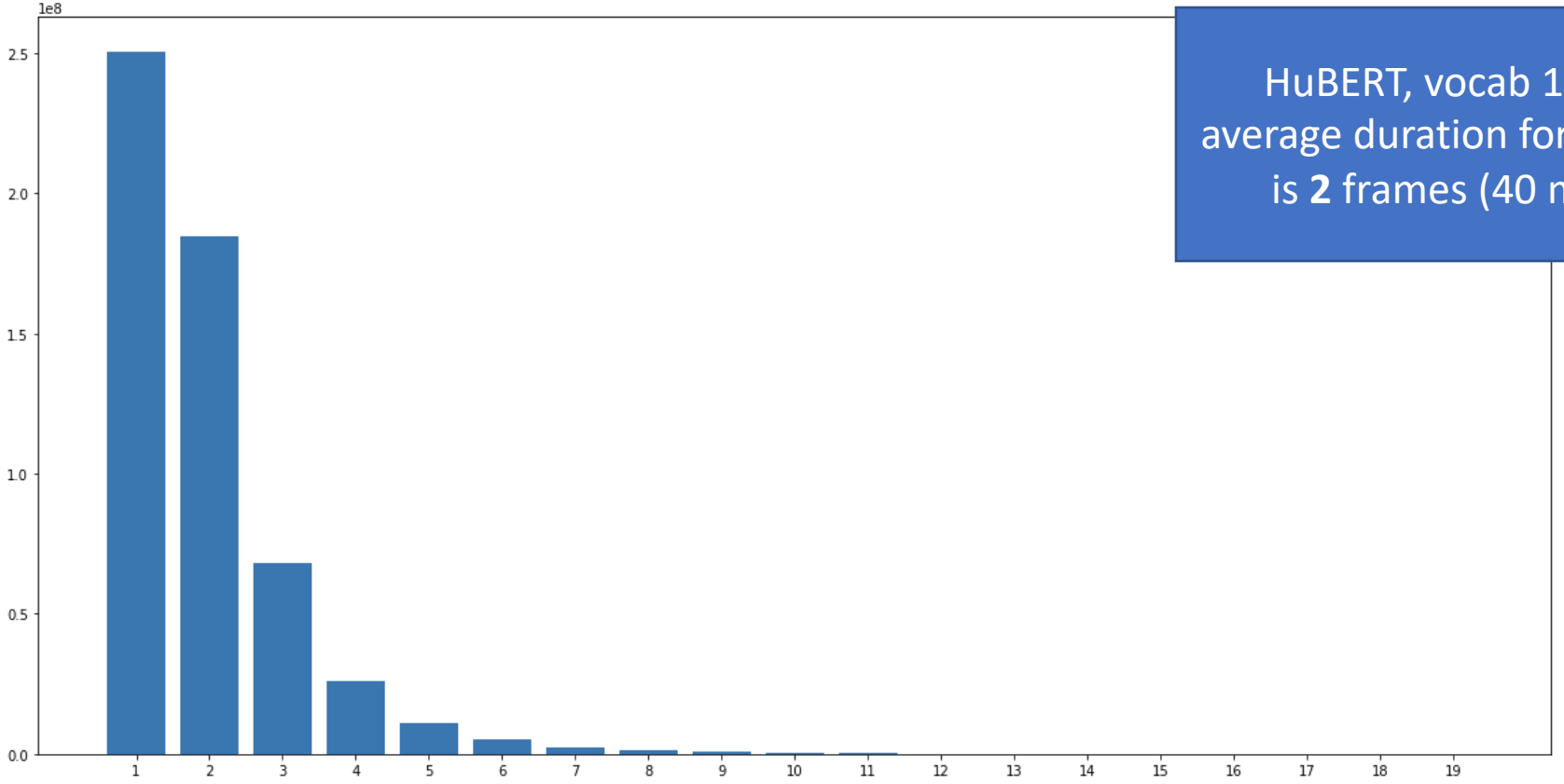
Unit Language Model (ULM)

Now we can run standard language modeling toolkit!

- Transformer-large LM from fairseq (Ott, 2019)
  - 12 layers, 16 heads, …
- Standard LM training recipe, copied from the fairseq LM page
- A “better” 6K hours part (Rivière & Dupoux, 2021) of 60K Libri-Light (Kahn et al, 2020)

Unit duplication

- Speech is continuous: the same unit can be repeated for a few times

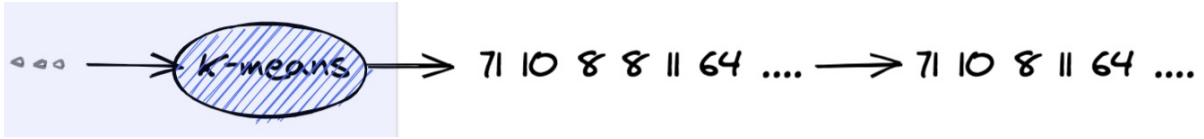


HuBERT, vocab 100: average duration for a unit is **2** frames (40 ms)

Unit duplication

Observation: Duration of a unit might be not so important for “content” modelling

- Let’s throw away the repetitions

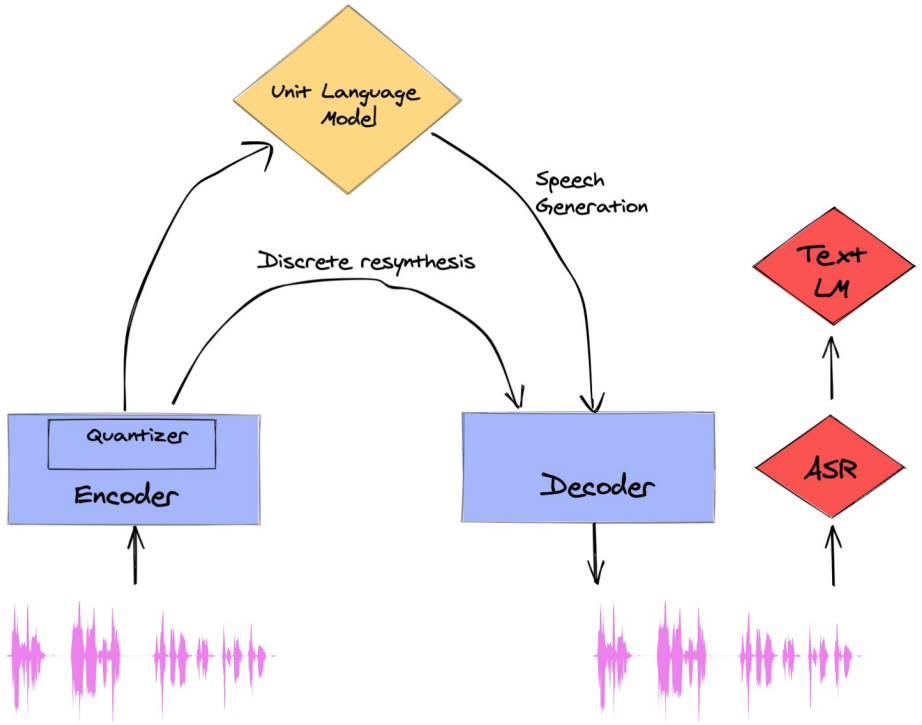


- Proved to be very useful
  - saves limited attention span

Metrics

Metrics: what is *good*?

	Representation		Generation		
	Task	**Metric**	**Task**	**Metric**	**Human evaluation**
Language Model	Spoken LM	Lexical & Grammatical judgements	Speech Generation	PPX@o-VERT, VERT@o-PPX, AUC, Continuation-BLEU	MMOS
Acoustic Model	Unit discovery	ABX, bitrate	Speech Resynthesis	ASR PER/CER	CER, MOS



Unit Discovery: ABX

- Zero-shot metric measuring the quality of learned representations
- We consider triplets (A,B,X) of sequences that can only differ in a central phoneme:
  - A: bit, B: bet, X: bit
  - We measure if dist(A, X) < dist(B, X)
  - Aggregate error rate
  - (X can come from the same or different speaker)

The Zero Resource Speech Challenge 2015: Proposed Challenges and Results. Versteegh et al., 2015

Unit Discovery: bitrate

- Bits/s required to encode the unit stream, assuming a unigram model

$$B(U) = -\frac{n}{D} \sum_i p(u_i) \log_2 p(u_i)$$

Units per second

Entropy

The Zero Resource Speech Challenge 2019: TTS Without T. Dunbar et al, 2019.

Spoken Language Modelling

- **sWUGGY:** measures the lexical ”knowledge” of the model
  - 5,000 similar-sounding pairs of word and non-words
  - “brick” vs. “blick”
  - Score the pairs using the trained ULM
  - Accuracy of LM-prob(word) > LM-prob(non-word)

The Zero Resource Speech Challenge 2021: Spoken language modelling. Dunbar et al, 2021.

Spoken Language Modelling

- **sBLIMP:** measures the syntactical ”knowledge” of the model
  - 6,300 pairs of syntactically correct and incorrect utterances
  - “the dogs sleep” vs. “the dog sleep”
  - Score the pairs using the trained ULM
  - Accuracy of LM-prob(correct) > LM-prob(incorrect)

The Zero Resource Speech Challenge 2021: Spoken language modelling. Dunbar et al, 2021.

Spoken Language Generation

”Convert” generated unit-utterances into wave and then into English texts with an off-the-shelf ASR

- (wav2vec 2.0-based, Baevski et al., 2020)

- Now we have a model that generates texts
- Evaluation reduces to evaluating the “quality” of the generated utterances

Spoken Language Generation (PPX)

- Median Perplexity:
  - Measure PPX according to an off-the-shelf large text LM
  - Large text Transformer LM, trained on NewsCrawl (Ng et al., 2019)

Spoken Language Generation (failure modes)

- ULM can produce few low-perplexity utterances
  - THIS IS LIBRIVOX RECORDING ALL LIBRIVOX RECORDINGS ARE IN THE PUBLIC DOMAIN
- ULM generates utterances with repetitions within it
  - THE PROPERTY BY JAMES RESELL RED FOR LIBERATA OR BY JASON DOWNY THE PROPERTY BY JASON DOWNY THE PROPERTY THE PROPERTY THE PROPERTY THE PROPERTY

Spoken Language Generation (Diversity)

- Use **self-BLEU** (Zhu et al., 2018) to address the problem of corpus diversity.
  - Take one utterance from the corpus (ULM ⟼ TTS ⟼ ASR)
  - Treat all others as "reference translations”
  - Measure the highest similarity
  - Average over all utterances

Spoken Language Generation (Diversity)

- We introduce auto-BLEU to address the problem of repetitions within an utterance.

$$\texttt{auto-BLEU}(u,k) = \frac{\sum_s \mathbb{1}\left[s \in (NG_k(u) \backslash s)\right]}{|NG_k(n)|}$$

- Putting both diVERsiTy metrics together.
  - VERT = Geometric Mean(self-BLEU, auto-BLEU)

Spoken Language Generation: Protocol

Two evaluation scenario:

- Unconditional generation
- Prompted generation
  - conditioned on 3 sec prompt from LibriSpeech test-clean

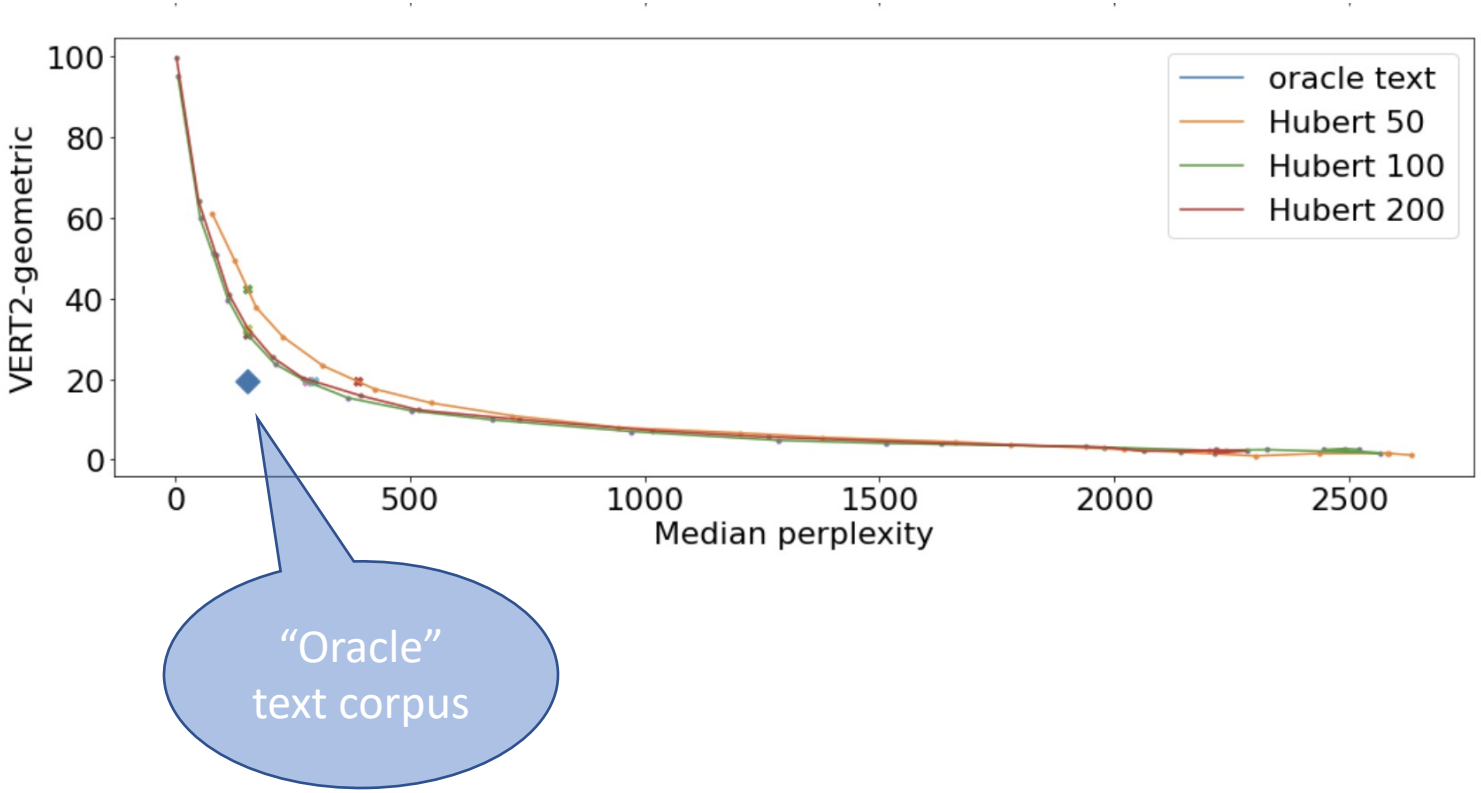
In both cases, we control the durations to be the same as in LibriSpeech test-clean – so that we can compare metrics to ones of the text

Spoken Language Generation: Sampling

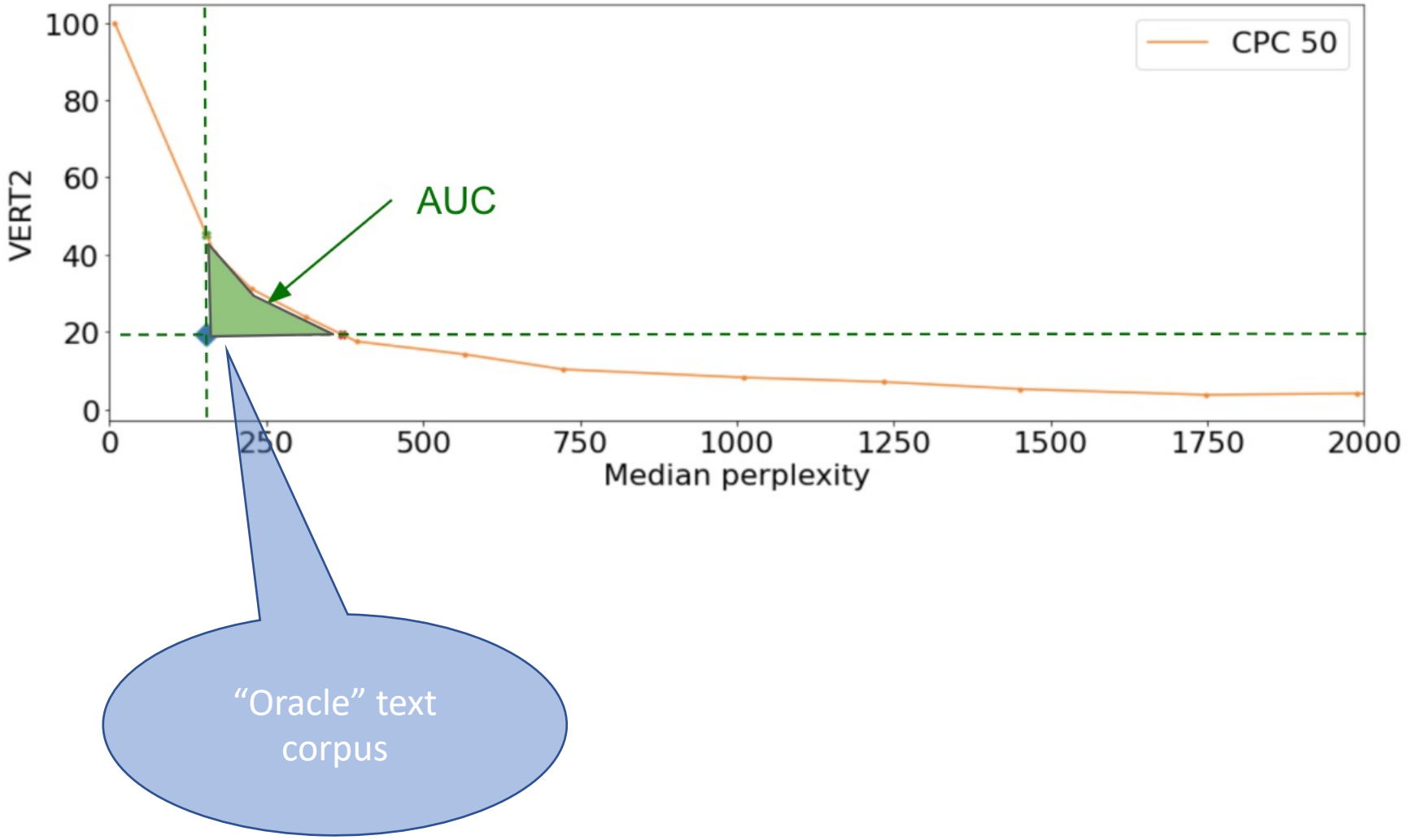
- Sampling with temperature
- Problem: temperature means different things for different vocabulary sizes and unit systems

Spoken Language Generation (PPX vs. Diversity)

We vary T in {0, 0.2, … 3.0}



Spoken Language Generation: PPX vs Diversity



Spoken Language Generation: at what temperature?

What if we had to select one “best” temperature? For instance, for running a human subjective evaluation?

- Continuation-BLEU:
  - Take a 3s prompt from each utterance in a hold-out dataset
  - Sample k (=10) possible continuations
  - Measure the text-level BLEU (after ASR) similarity between the ground-truth and generated samples
- Select the temperature with highest Continuation-BLEU

Spoken Language Generation: Human Evaluation

- Meaningfulness-MOS (MMOS): judges were asked to evaluate how natural (considering both grammar and meaning) the content of a given recording is

- For each (representation type × number of units)
  - 100 samples
  - at temperature that maximises Cont-BLEU2

Results

Comparing representations & vocabulary sizes

Compared Systems

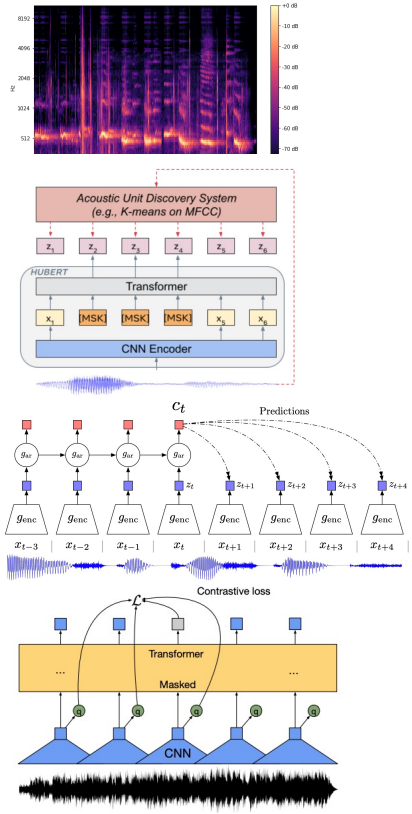
- Topline: ASR + TTS (wav2vec-2.0, Tacotron2)
- Baseline: LogMel

- HuBERT (Hsu, 2021)

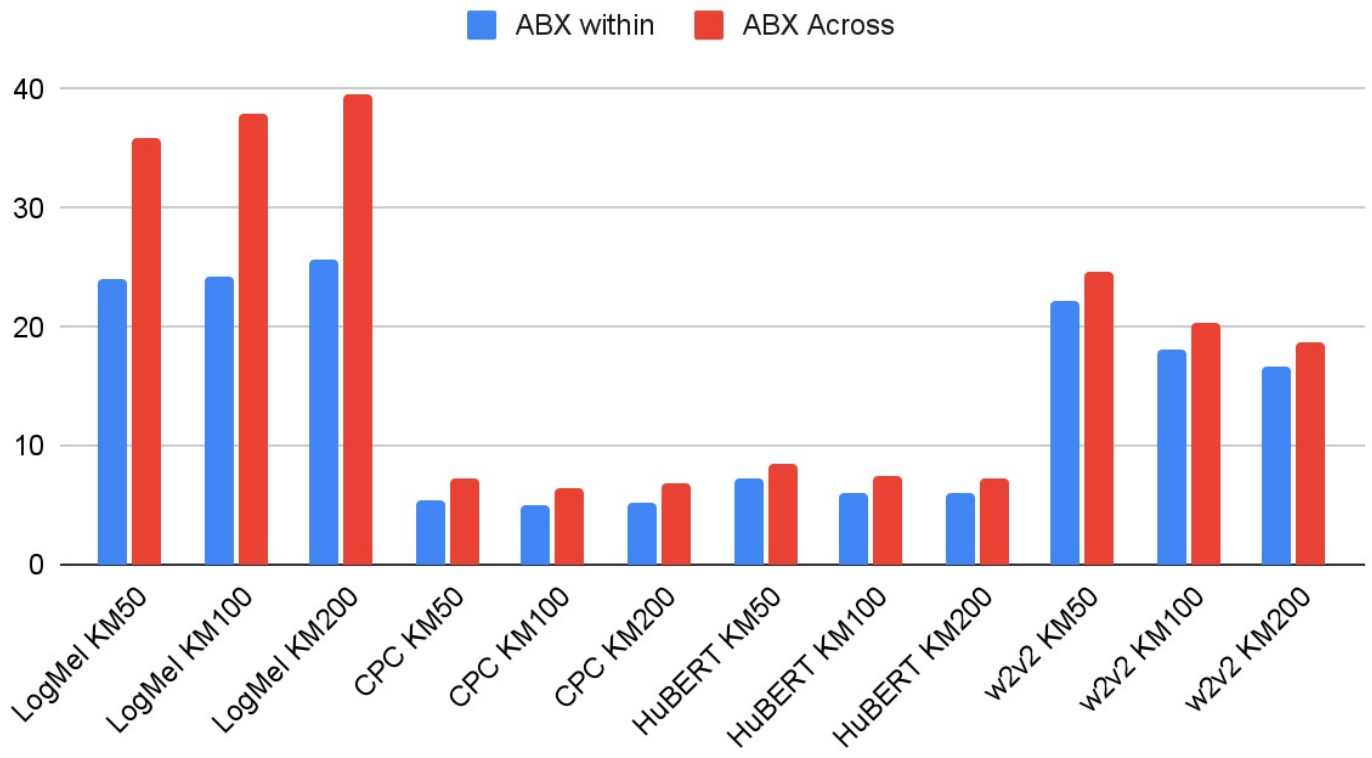
- CPC (van den Oord, 2019)

- wav2vec-2.0 (Baevski, 2020)

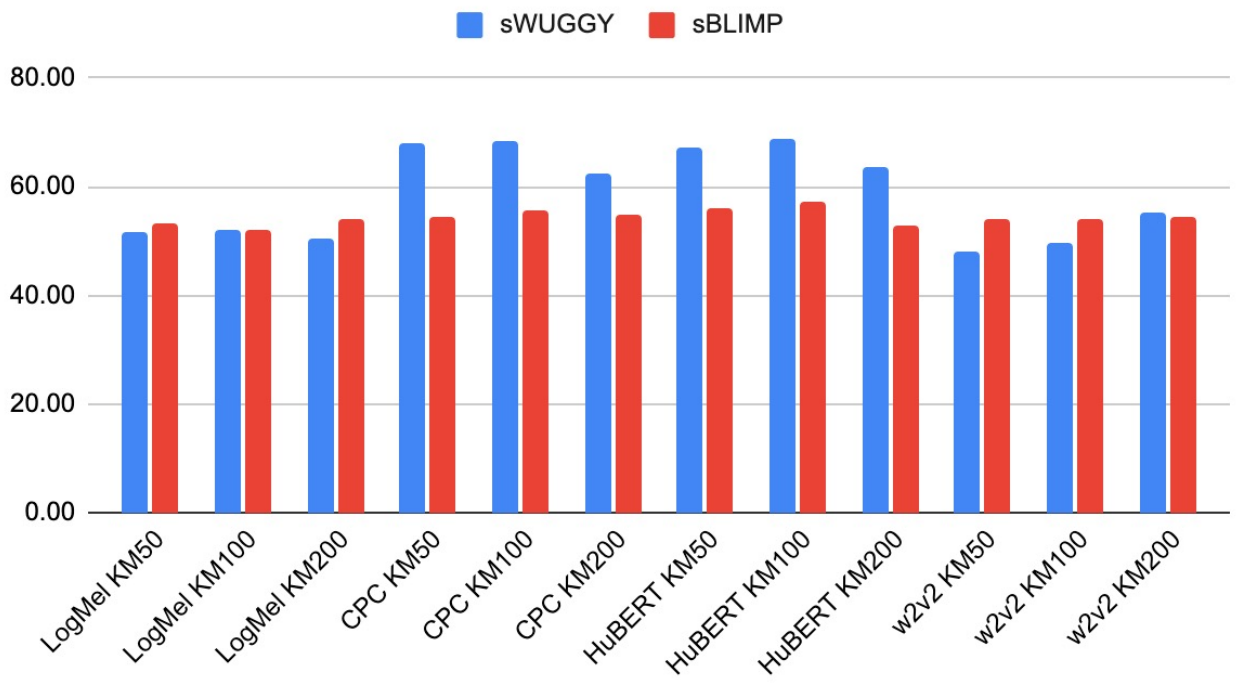
Vocabularies of 50, 100, 200 units



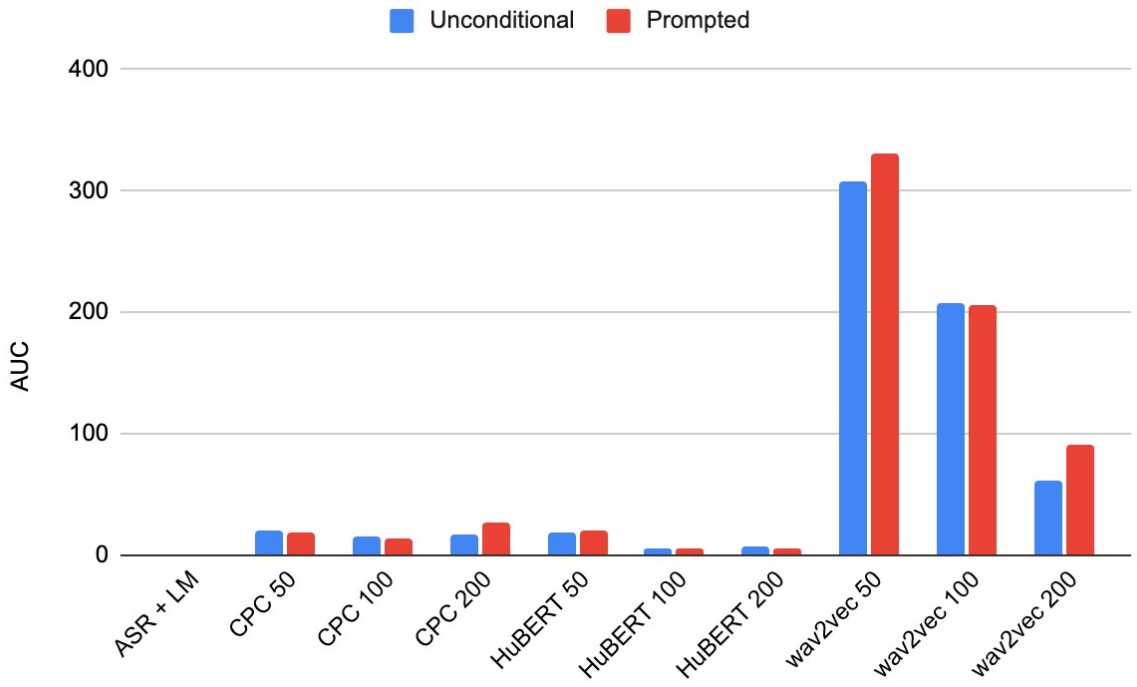
Unit Discovery: ABX

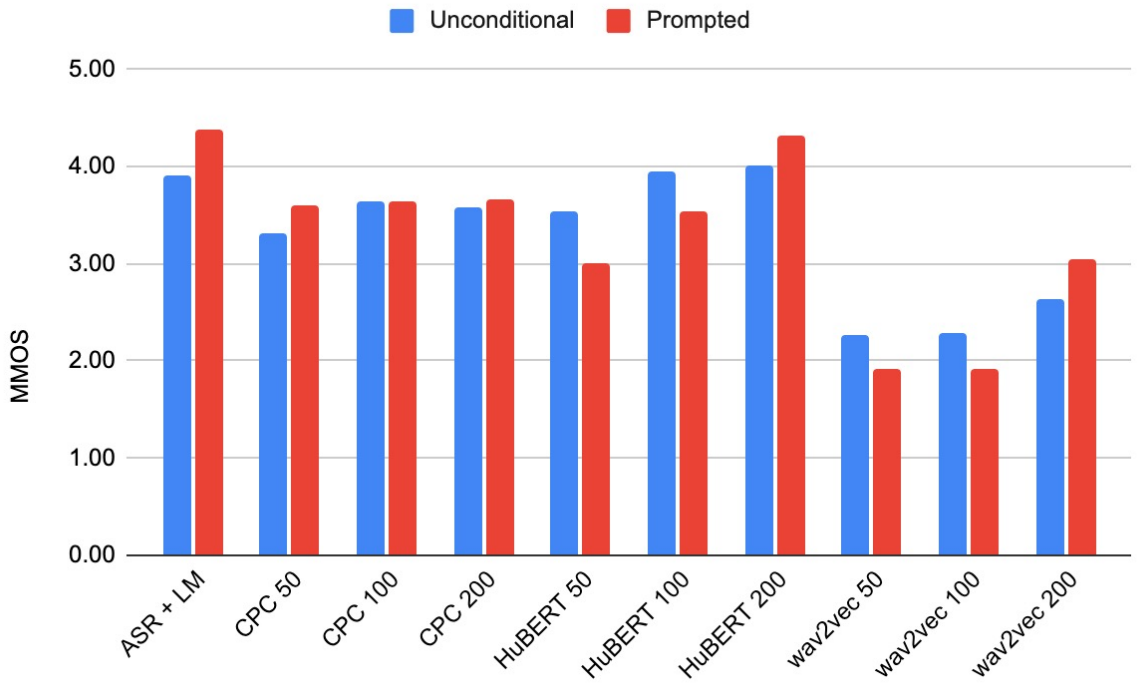


Spoken Language Modelling: sWUGGY & sBLIMP

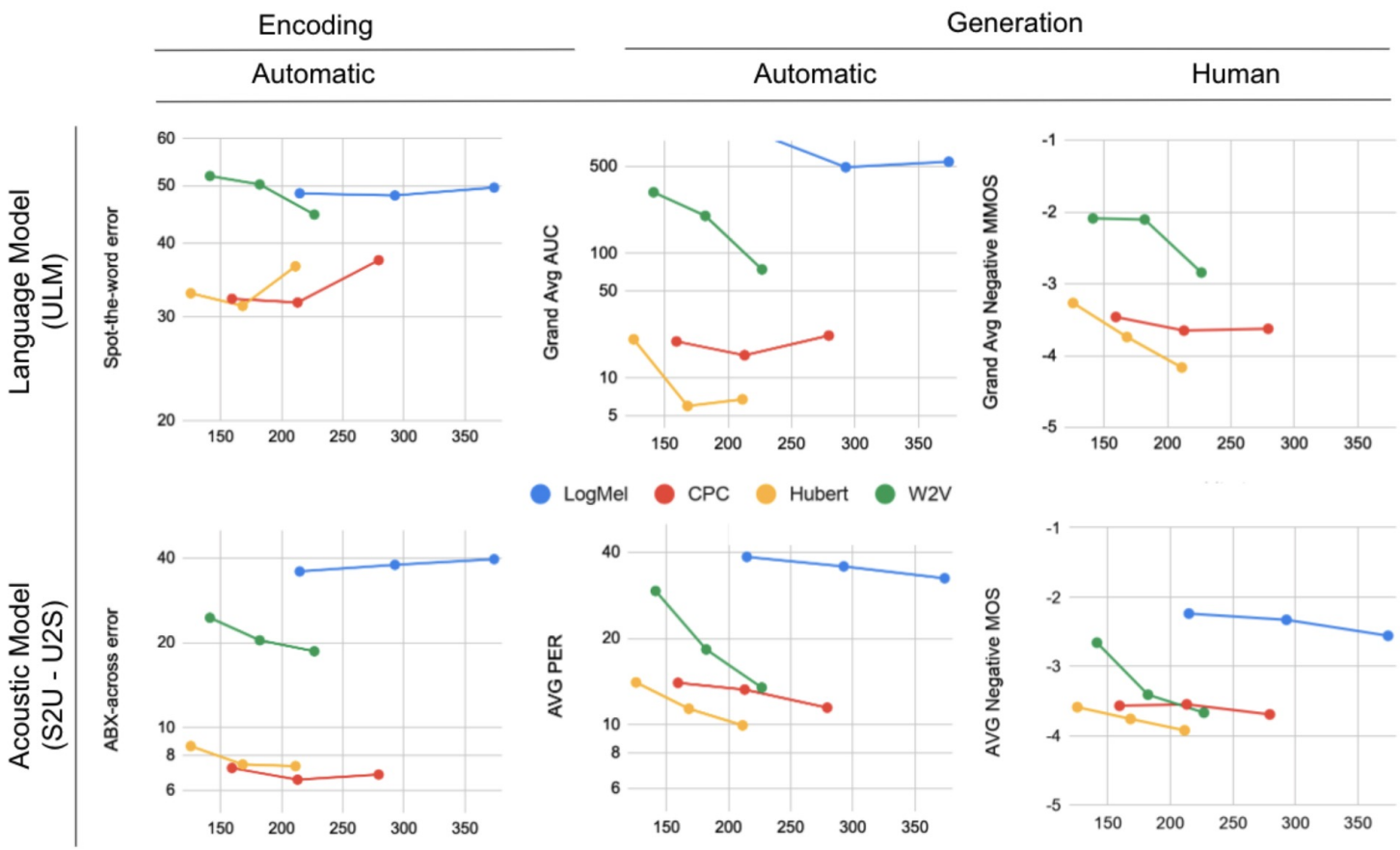


Spoken Language Generation





Aggregated metrics vs. bitrate



Interim conclusions

- We introduced a new setup/task, Speech Generation that allows to produce new speech utterances in an unsupervised way
- A mix of: acoustic unit discovery, spoken language modelling, discrete speech resynthesis, and text generation
- A suite of end-to-end metrics well correlated to human judgements
- Allows using generic NLP toolset on speech

What is missing?

- Speech is so much more than just “textual” content: there is expressive/prosodic information that is very valuable for humans

What is missing?

Questions:

- Can we improve content modelling by jointly modelling prosodic information with “content” units?
- Can we generate consistent & diverse prosody in Speech Generation?

Prosody = pitch (F0) + tempo

Text-Free Prosody-Aware Generative Spoken Language Modeling

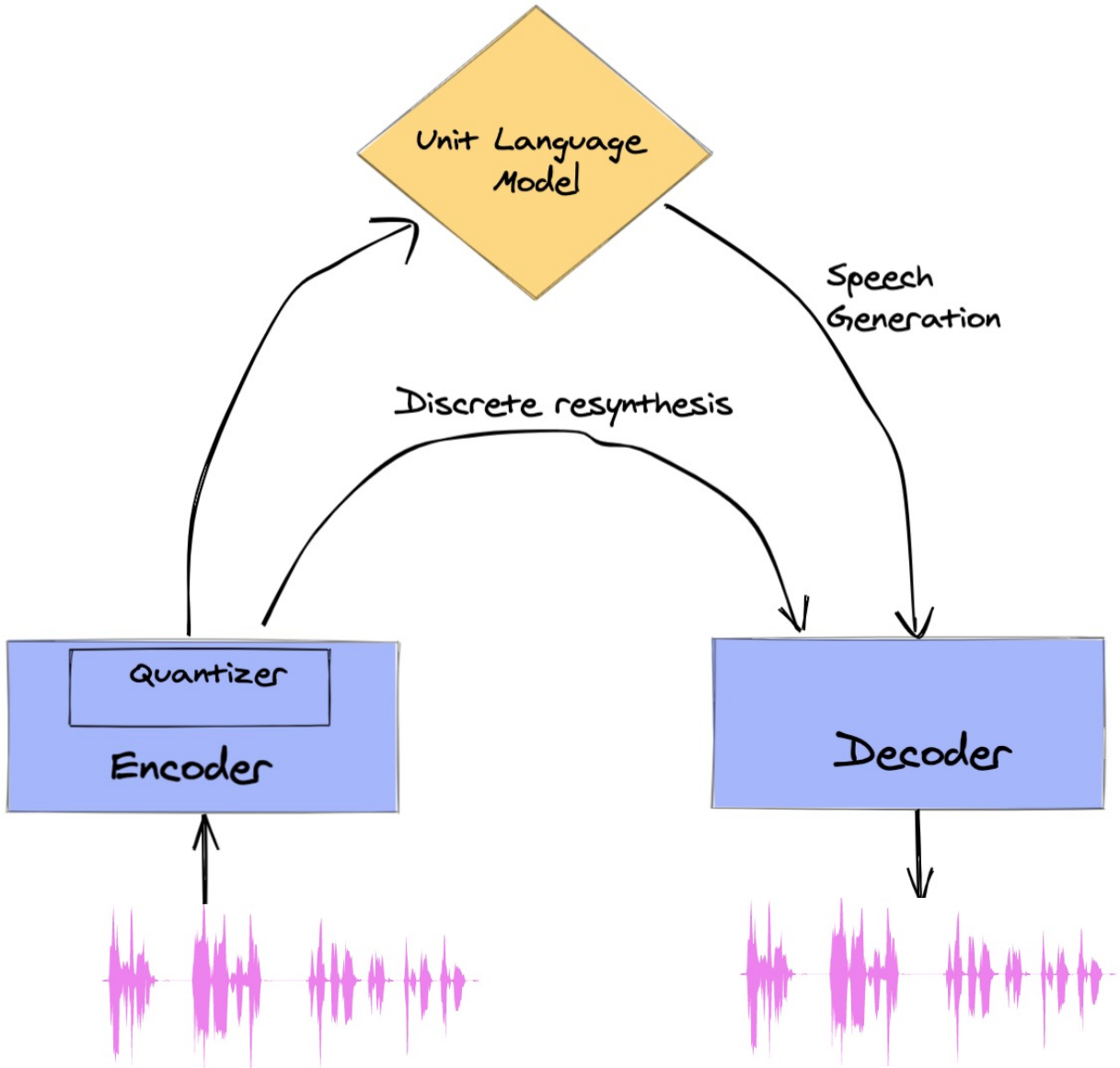
Eugene Kharitonov*, Ann Lee*, Adam Polyak, Yossi Adi,
Jade Copet, Kushal Lakhotia, Tu-Anh Nguyen, Morgane Rivière,
Abdelrahman Mohamed, Emmanuel Dupoux, Wei-Ning Hsu
Facebook AI Research

Examples

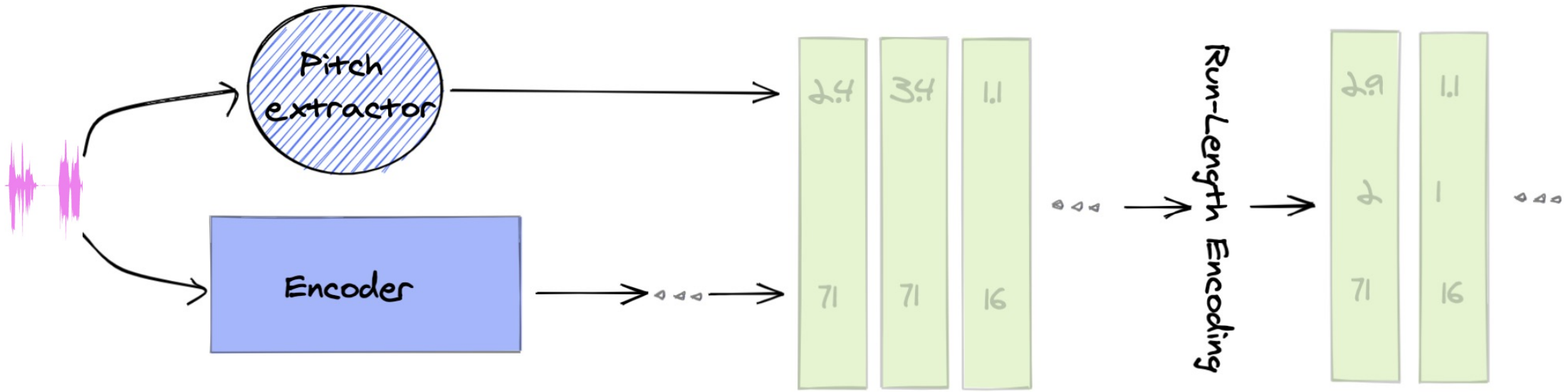
- Transformer-based language model, trained on audio w/o text or labels
- We prompt with 3s audio from the test set (or even a different dataset!)
- The model auto-regressively generates the continuation:
  - Content,
  - Pitch,
  - Duration

speechbot.github.io/pgslm/

Architecture changes


Unit Language Model
Speech Generation
Discrete resynthesis
Quantizer
Encoder
Decoder

Multi-stream speech representation



Data representation

Speech is represented as three sync streams

- Unit stream:
  - 100 centroids, HuBERT-base (Hsu et al., 2020), 50 frames per second
- Duration stream:
  - Duration of the corresponding token in frames
- F0 stream:
  - speaker-mean normalized log-F0
  - averaged across voiced frames in the segment

Prosody representation

We experiment with two ways of encoding prosodic streams:

- Continuous: duration and pitch F0 are continuously valued variables
- Discrete:
  - duration is capped and quantized
  - F0 is quantized in equiprobable bins (32 in reported experiments)

Loss

Unit stream

F0 stream

$$L(p(u_t, d_t, f_t), u_t^*, d_t^*, f_t^*) = L_u(p(u_t), u_t^*) + \alpha \cdot L_d(p(d_t), d_t^*) + \beta \cdot L_f(p(f_t), f_t^*)$$

Duration

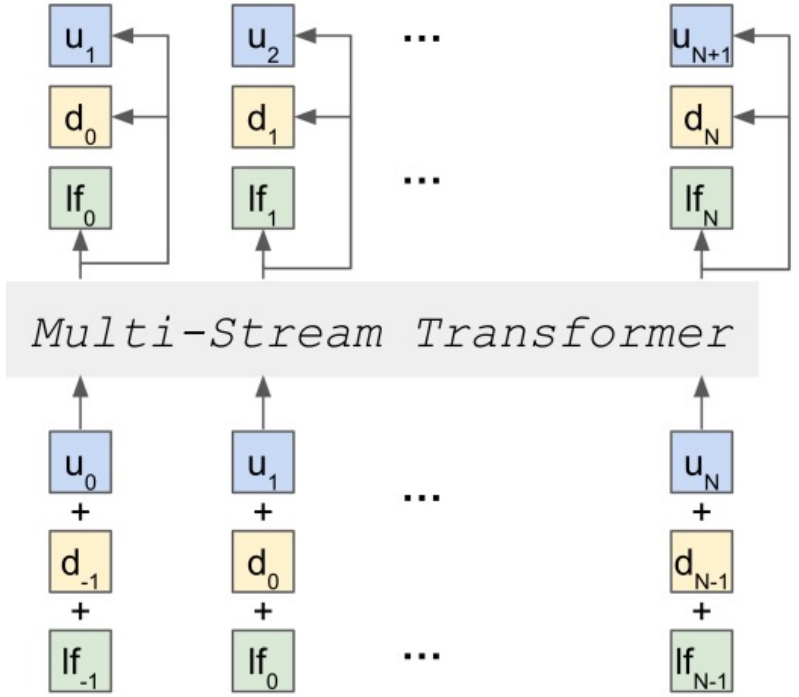
- Sum of per-stream losses
- Found the results to be robust w.r.t. the relative weights, so set them to 0.5
- NLL for discrete- and L1 (MAE) if continuous-valued streams

Sampling

- Discrete
  - Sampling with temperature from the multinomial distribution
  - Temperature is tuned on dev set
- Continuous
  - Treat as a (truncated) Laplace distribution
  - Duration: round to the nearest natural number
  - Scale parameter (the counterpart of the temperature) is tuned on the dev set

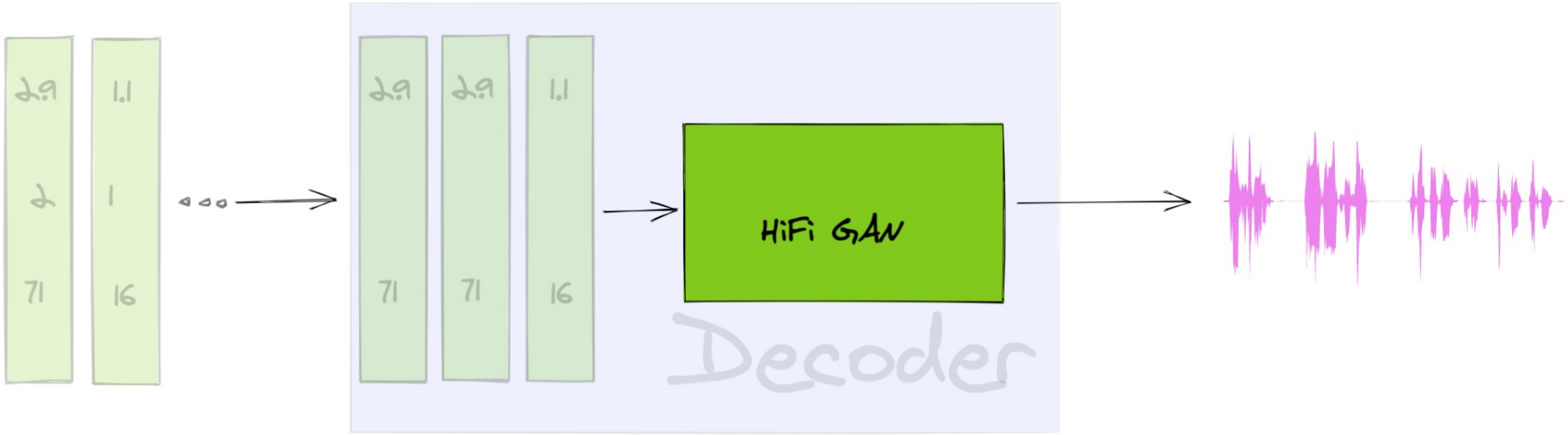
Multi-Stream Transformer

- We found that it can be useful to delay prosodic streams w.r.t. the unit stream
- Prosody is predicted when the token is already known – better prosody modelling
- On the other hand, token prediction has less prosody context – a bit worse unit modelling



Decoder

- HiFi GAN: better quality / faster generation
- Unroll segments back into frames, using average log-normalized F0
- Trained on the Blizzard dataset (SynSIG, 2013)



We use HiFi GAN (Kong et al., 2020) variant by Polyak et al. (2021)

Metrics

Metrics

Questions:

- Can we improve content modelling by jointly modelling content & prosodic information?
- Can we generate consistent & diverse prosody?

Metrics

- Teacher-forced metrics
- Generation metrics
- Human-based evaluation

Teacher-forced

- Can we improve content modelling by jointly modelling content & prosodic information?
  - NLL on the hold-out data (given all the historical context for a frame)

- Can we generate consistent prosody?
  - L1 loss on prosodic streams (given all the historical context for a frame)
  - De-quantize if discrete

Speech Generation

- Can we improve content modelling by jointly modelling prosodic information alongside with “context” units?
  - Continuation-BLEU: 20 continuations of the same duration as ground-truth
  - Turn it into text with the same off-the-shelf ASR
  - Maximal word-level BLEU similarity between actual and generated continuation

Speech Generation

- Can we generate consistent prosody?
  - Continuation-L1: loss on continuation after a prompt (min over 20 samples, Min-MAE)
  - Correlation between prompt's per-frame-mean (pitch, duration) and the generated per-frame-mean (pitch, duration)
- Can we generate diverse prosody?
  - Standard deviation of generated prosodic streams

Metrics: human evaluation

- Human MMOS:
  - sound quality,
  - meaningfulness (how natural the text content is considering both grammar and meaning),
  - prosody (how consistent and natural the intonation and the rhythm are)

Data and Hyperparameters

Data:

- LibriSpeech 960h (Panayotov et al., 2015)
- “better” subset 6K h (Rivière and Dupoux, 2020) of Libri-Light (Kahn et al., 2020)

Standard architecture sizes from fairseq:

- Base (6 layers, 8 heads, 512/2048 embedding/FFN size)
- Large (12 layers, 16 heads, 1024/4096 embedding/FFN size)

Adam w/ standard optimization tricks

Results

Prosodic Inputs Are Useful for Content Modelling

Input	Output	u NLL↓
Base MS-TLM, HuBERT units, trained on LS960		
u	u	1.522
(u, d, f)	u	**1.336**

Prosodic Inputs Are Useful for Predicting Prosody

Input	Output	Quantized?	d MAE↓	f MAE↓
Large MS-TLM, HuBERT units, trained on LL6k				
u	(u, d, f)		0.563	0.095
(u, d, f)	(u, d, f)		**0.527**	**0.043**
u	(u, d, f)	✓	0.586	0.116
(u, d, f)	(u, d, f)	✓	**0.543**	**0.047**
Base MS-TLM, CPC units, trained on LS960				
u	(u, d, f)	✓	1.302	0.122
(u, d, f)	(u, d, f)	✓	**1.181**	**0.045**
Base MS-TLM, Phone units, trained on LS960				
u	(u, d, f)	✓	2.748	0.150
(u, d, f)	(u, d, f)	✓	**2.419**	**0.079**

Speech Generation

Input	Output	Quant?	d			f			Max-Word-Cont-BLEU2 ↑
			min-MAE ↓	Corr.↑	Std.↑	min-MAE ↓	Corr.↑	Std.↑	
ground truth		n/a	.000	.463	1.32	.000	.520	.163	1.000
resynthesized		✓	.000	.464	1.32	.000	.315	.145	.943
Large MS-TLM, HuBERT units, trained on LL6k									
u	(u, d, f)		.542	.176	.942	.084	.093	.081	.488
u	(u, d, f)	✓	.542	.086	**.965**	.096	.217	.147	.489
(u, d, f)	(u, d, f)		.539	**.344**	.940	.081	**.494**	.076	.498
(u, d, f)	(u, d, f)	✓	**.536**	.242	.946	**.077**	.324	**.149**	**.499**

Human subjective evaluation

Input	Output	Quant?	Mean Opinion Score MOS	M-MOS	P-MOS
resynthesized			$3.21_{\pm 0.09}$	$3.95_{\pm 0.32}$	$3.87_{\pm 0.45}$
Large MS-TLM, HuBERT units, trained on LL6k					
u	(u, d, f)		$3.16_{\pm 0.19}$	$3.80_{\pm 0.25}$	$3.69_{\pm 0.42}$
u	(u, d, f)	✓	$2.66_{\pm 0.18}$	$3.36_{\pm 0.40}$	$3.15_{\pm 0.52}$
(u, d, f)	(u, d, f)		$3.31_{\pm 0.23}$	$3.76_{\pm 0.27}$	$3.78_{\pm 0.46}$
(u, d, f)	(u, d, f)	✓	$3.43_{\pm 0.20}$	$4.04_{\pm 0.20}$	$3.75_{\pm 0.48}$

Conclusion

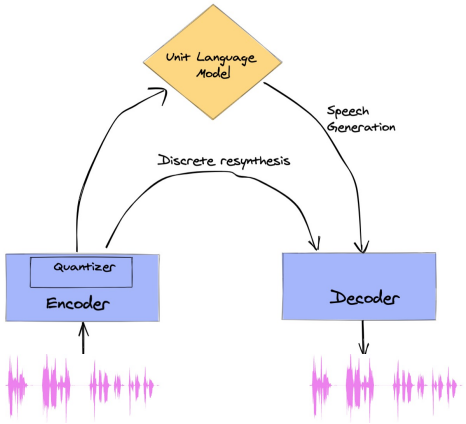
- We considered an extension of an earlier task, prosody-aware Speech Generation + a suite of metrics

- We showed that our multi-stream Transformer model can utilize prosodic inputs to improve content modelling
- Generates diverse and consistent prosody

We've discussed

- Spoken Language Modelling:
  - "Baseline" setup
    - How to evaluate its different components
    - Demonstrated agreement between proposed automatic metrics and human subjective judgements
  - Prosody-aware extension
    - How to incorporate prosodic information into an LM
    - Showed what benefits it brings about: (a) better content modelling, (b) consistent prosody continuation, (c) more diverse generated prosody

Thanks!



https://speechbot.github.io/