Textless NLP
towards language processing from raw audio

Emmanuel Dupoux
EHESS, META A
https://cognitive-ml.fr

LREC 2022, Marseille

What

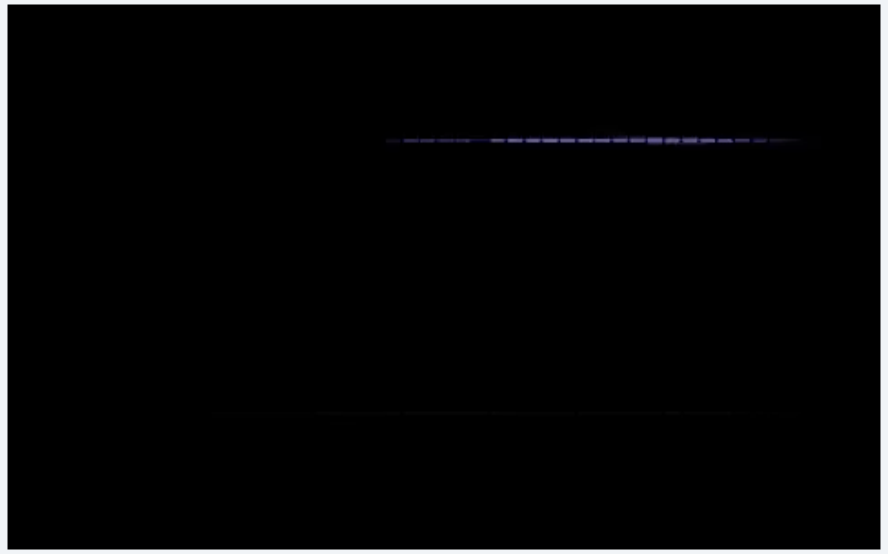
Textless NLP

Spoken language generation

Training AI models directly from raw audio recordings - no text or labels

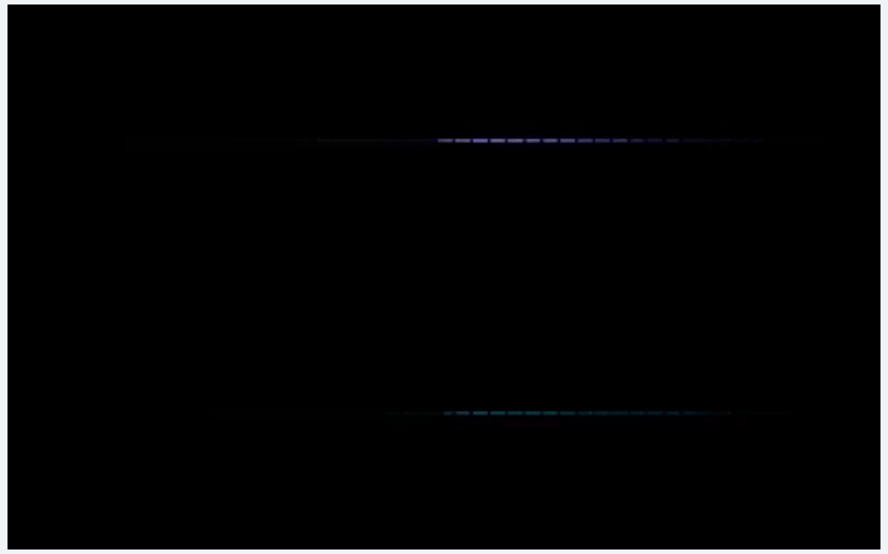
Spoken language is the primary means of human communication[1]

Yet, internet services are text based and struggle to capture nuances and richness of the oral modality.



Generating spoken dialogues with AI using our latest model[2]

Our agents reproduce naturalistic turn taking behavior including laughter and backchanneling, which is important for smooth human/agents interactions.

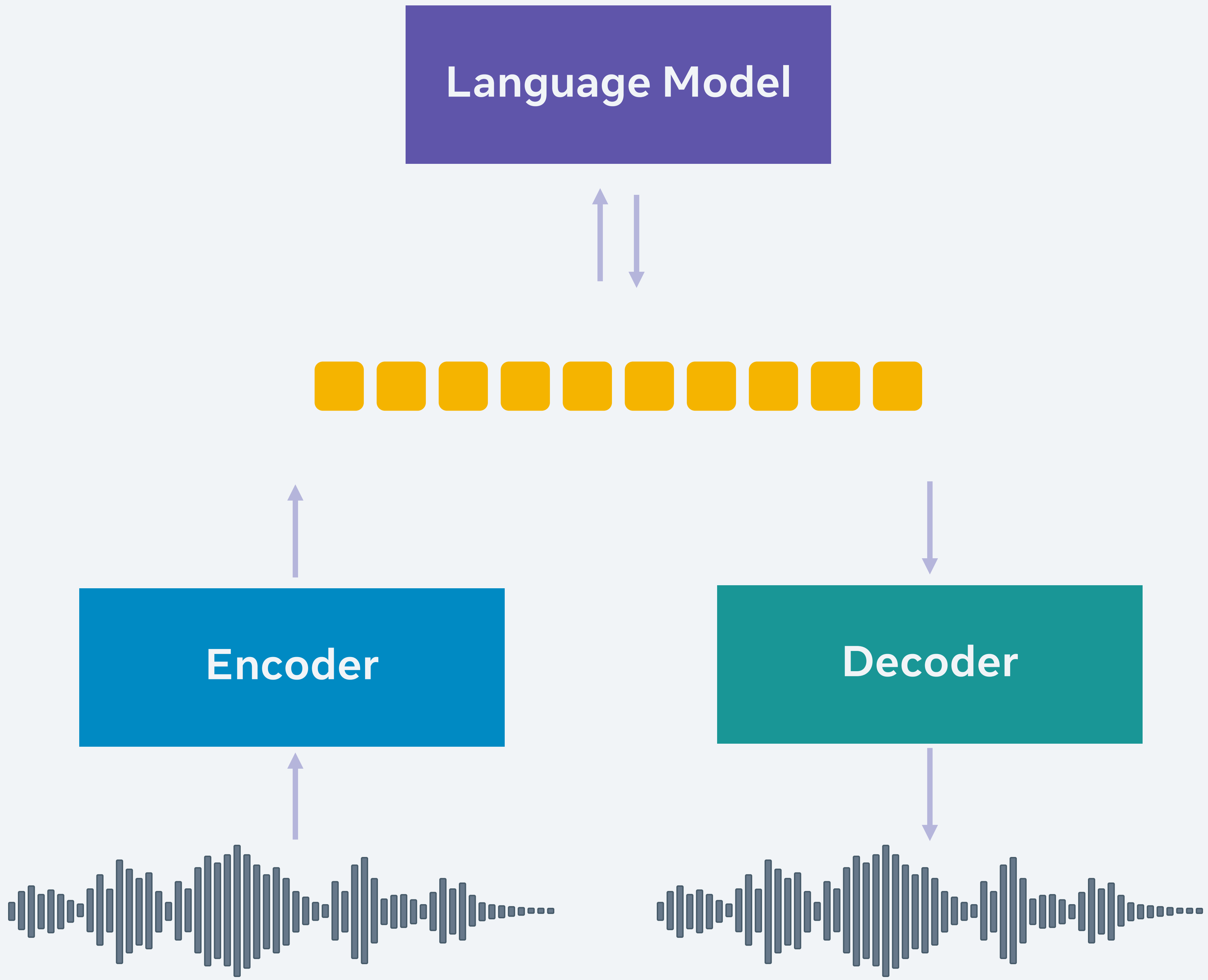


1. Fisher dataset
2. Nguyen et al. (2022)

How

Generative spoken language modeling

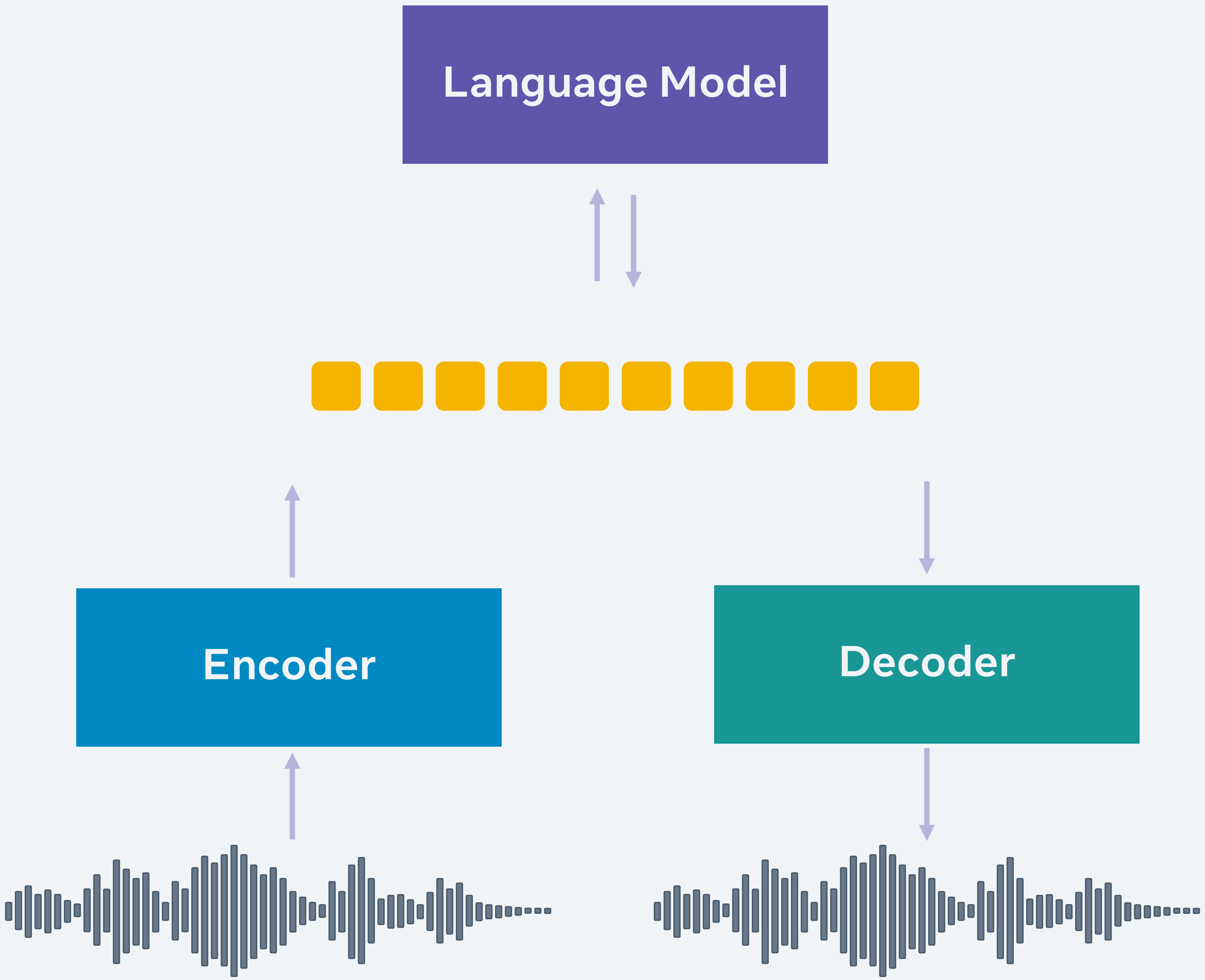
Self-supervised learning!



Generative spoken language modeling

Evaluation

Zero Resource Speech Challenge (ZRC) series

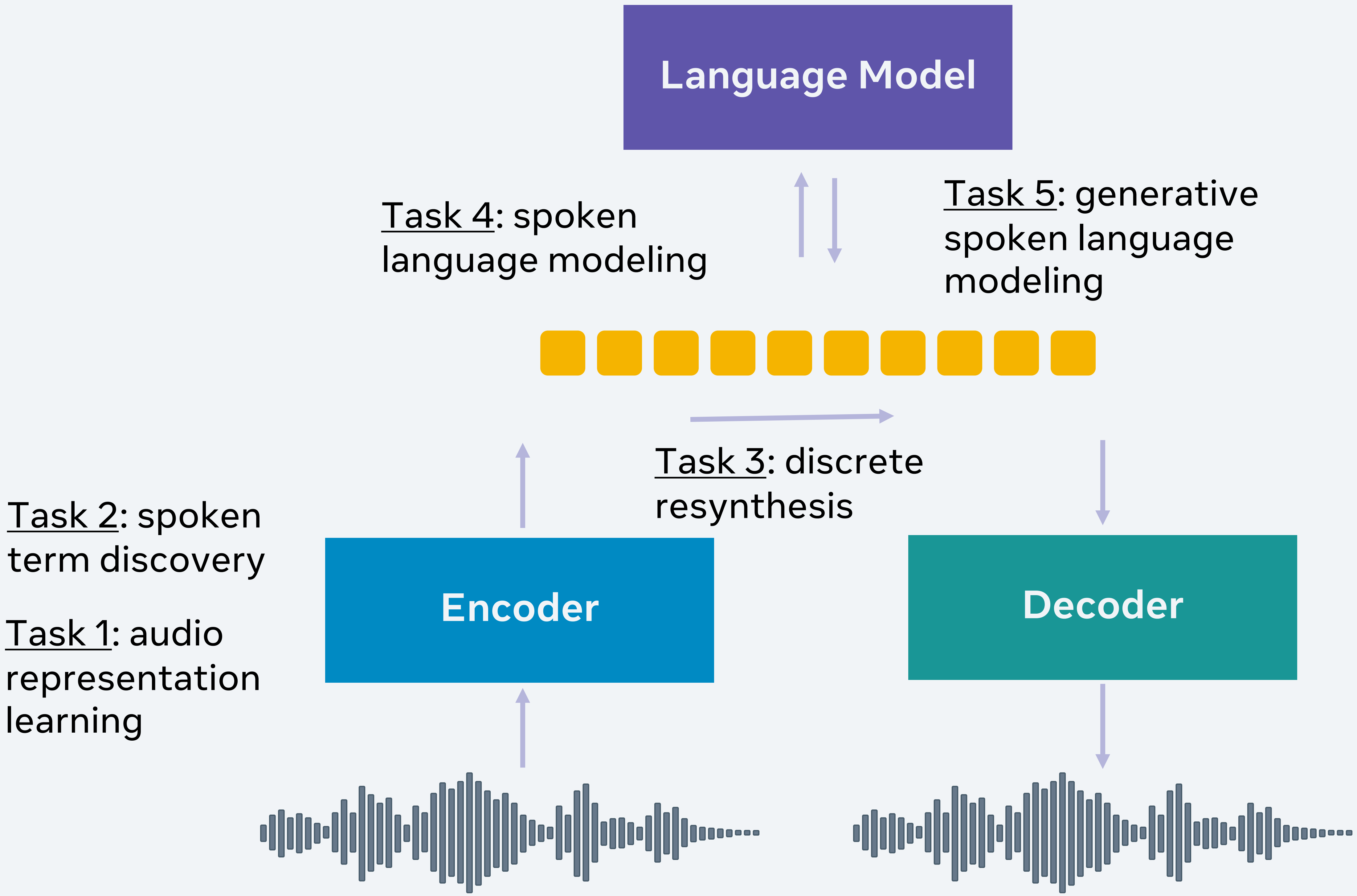


https://www.zerospeech.com

Generative spoken language modeling

Evaluation

Zero Resource Speech Challenge (ZRC) series



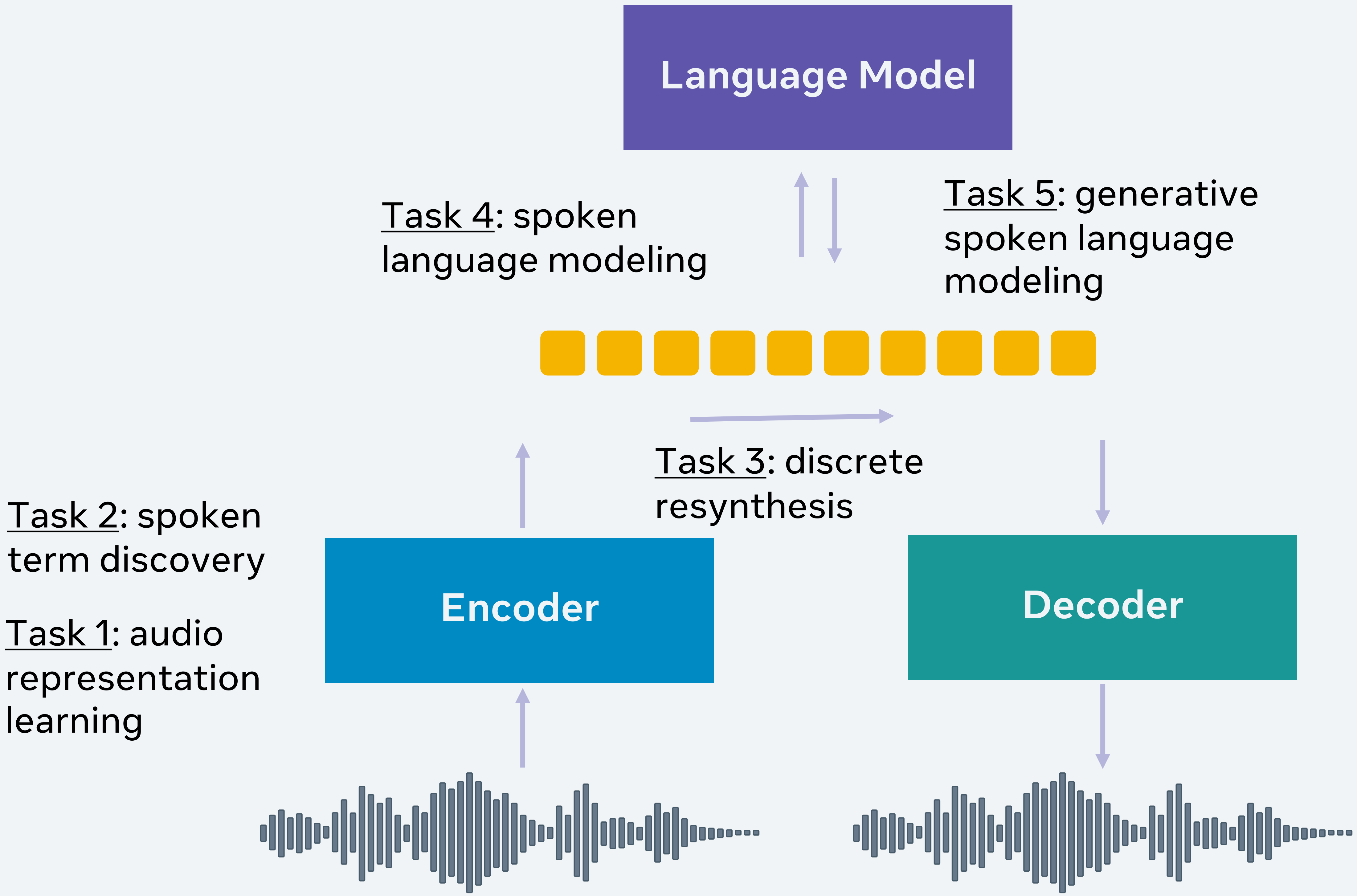
https://www.zerospeech.com

Generative spoken language modeling

Evaluation

Chall.	Tasks	Train Data
2015	T1, T2	English (Buckeye 5h), Xitsonga (2h30)
2017	T1, T2	English (45h), French (24h), Mandarin (2h30), German (25h), Wolof (10h)
2019	T3.	English (15h+4h40), Indonesian (15h+1h30)
2020	T1,T2,T3	reboot of ZR17, ZR19
2021a	T1,T4	English (Librispeech 960 or 100)
2021b	T1,T4	idem plus speech coco

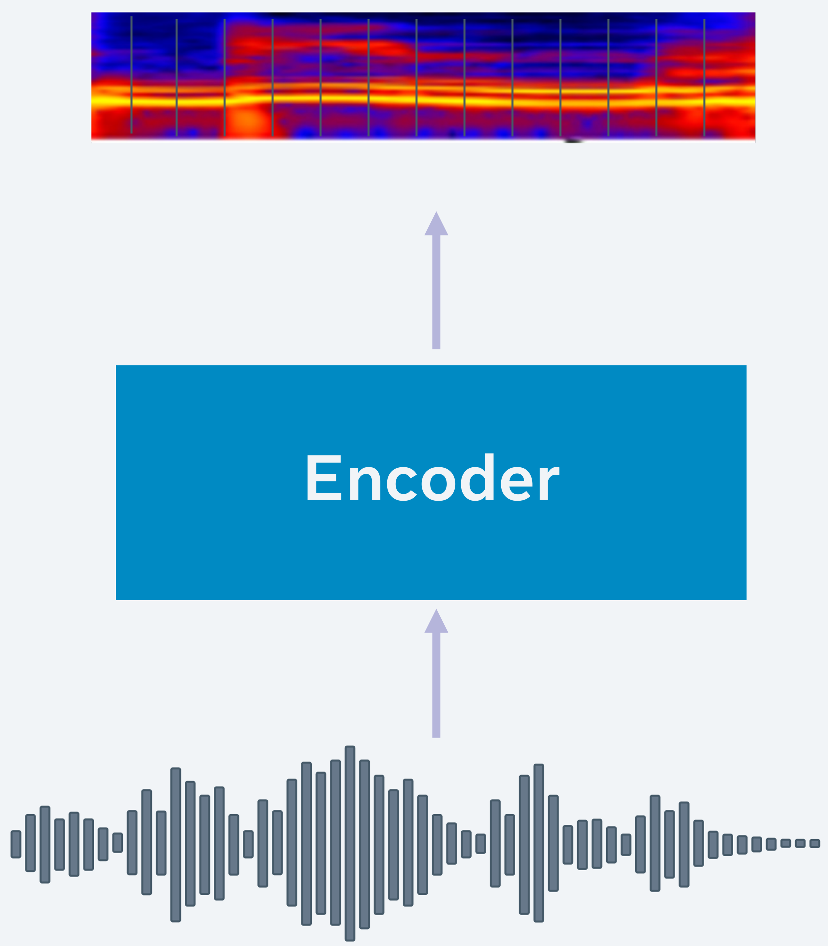
Zero Resource Speech Challenge (ZRC) series



https://www.zerospeech.com

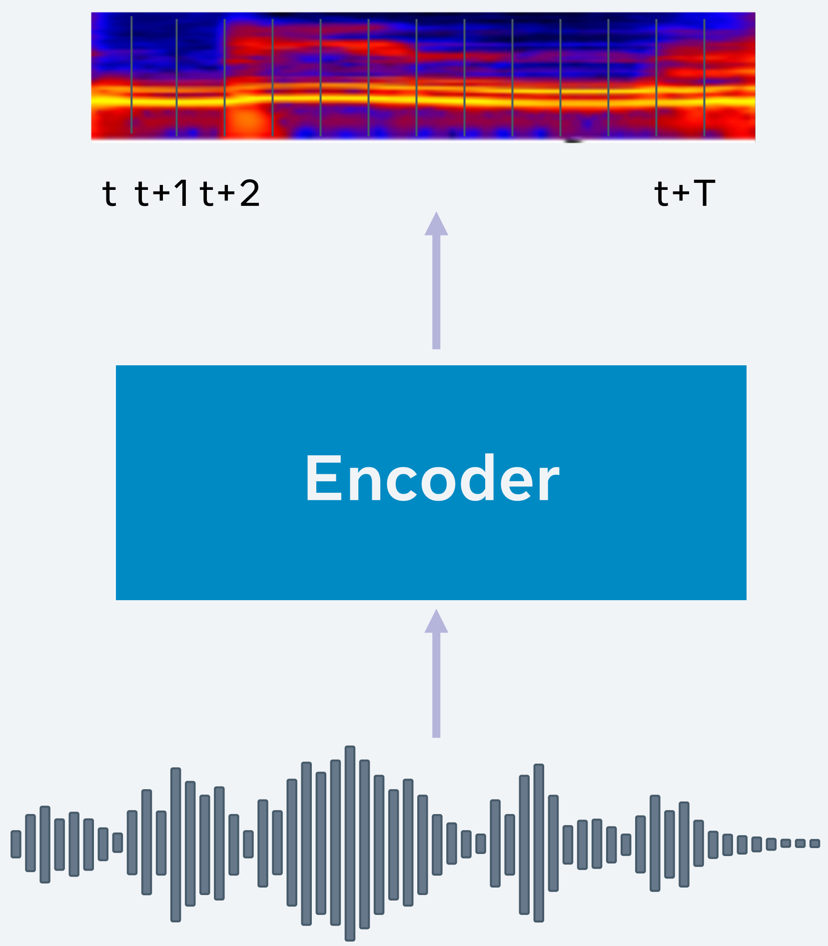
The encoder

Audio Representation Learning



The encoder

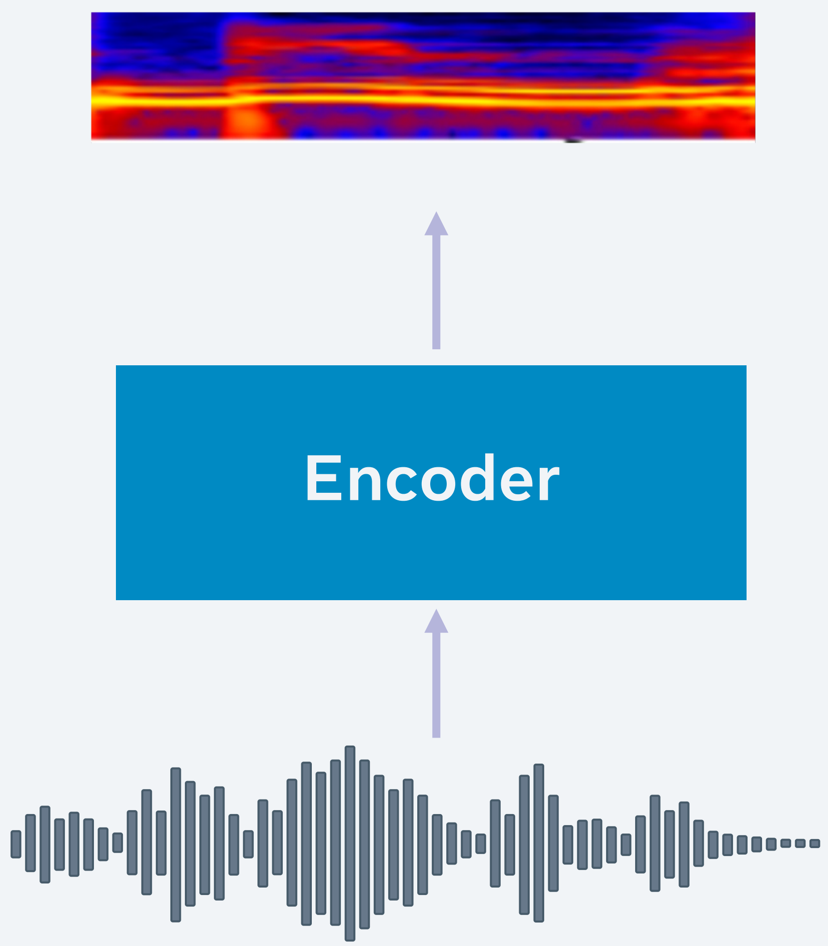
Audio Representation Learning



ZRC TASK 1:
Learning representations that encode linguistic information, and disregard non linguistic ones

The encoder

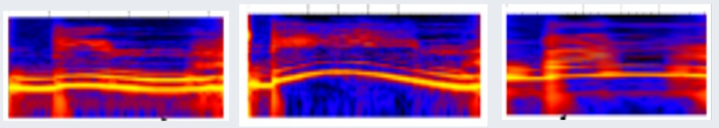
Audio Representation Learning



ZRC TASK 1:
Learning representations that encode linguistic information, and disregard non linguistic ones

<u>Evaluation:</u> *ABX discrimination*

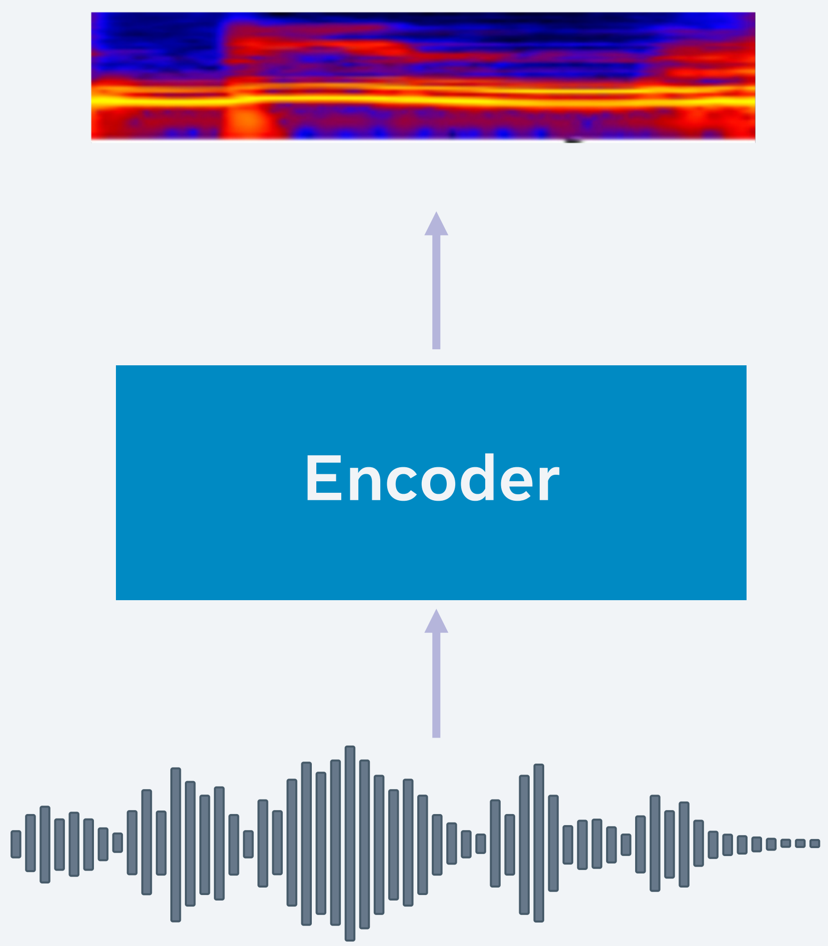
a	b	x
bit_{T1}	bet_{T1}	bit_{T2}



d(a,x) < d(b,x) ?

The encoder

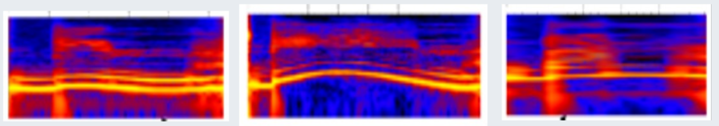
Audio Representation Learning



ZRC TASK 1:
Learning representations that encode linguistic information, and disregard non linguistic ones

<u>Evaluation:</u> *ABX discrimination*

a	b	x
bit_{T1}	bet_{T1}	bit_{T2}



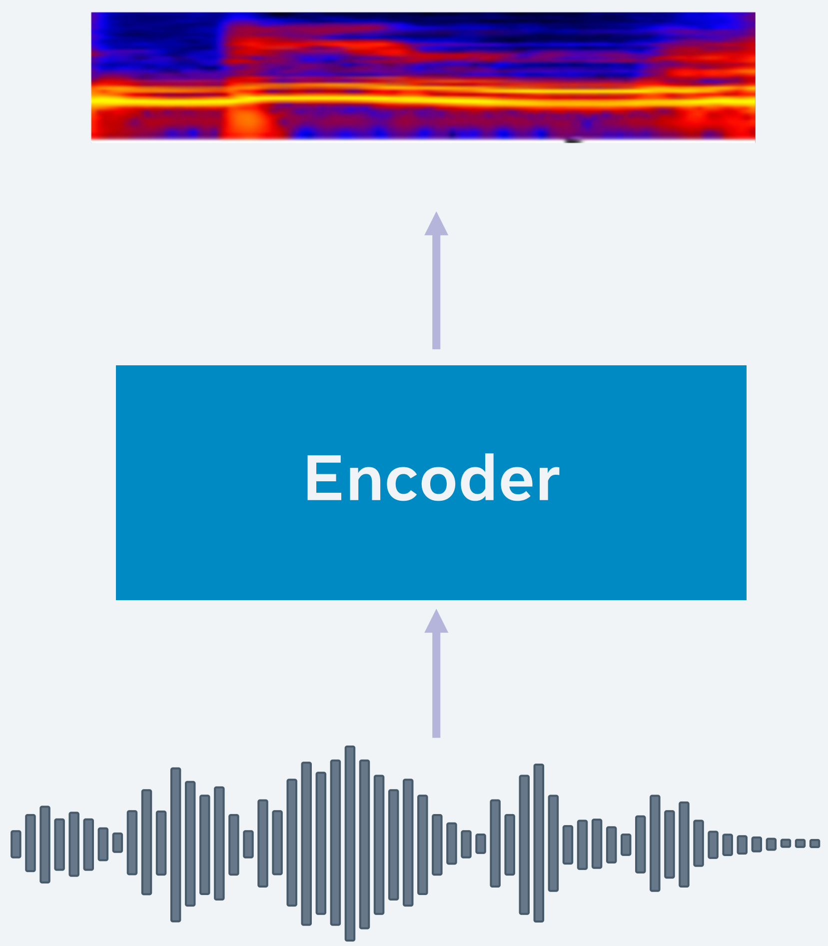
$d(a,x) < d(b,x)$?

<u>Main idea:</u> *information compression*

- Spectral information (MFCC):
  20kbit/sec
- Telephone, speech codec:
  8kbit/sec (2.5x reduction)
- Text (phonemes):
  50bits/sec (***400x reduction !)***

The encoder

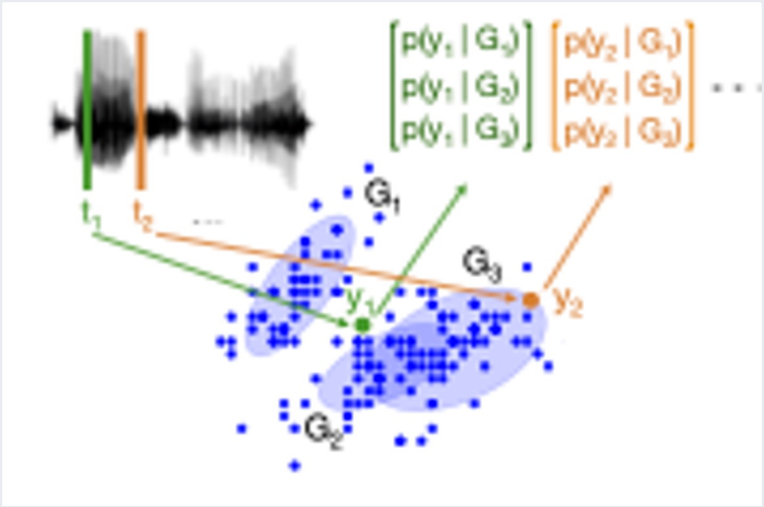
Audio Representation Learning



Best models

Compressive

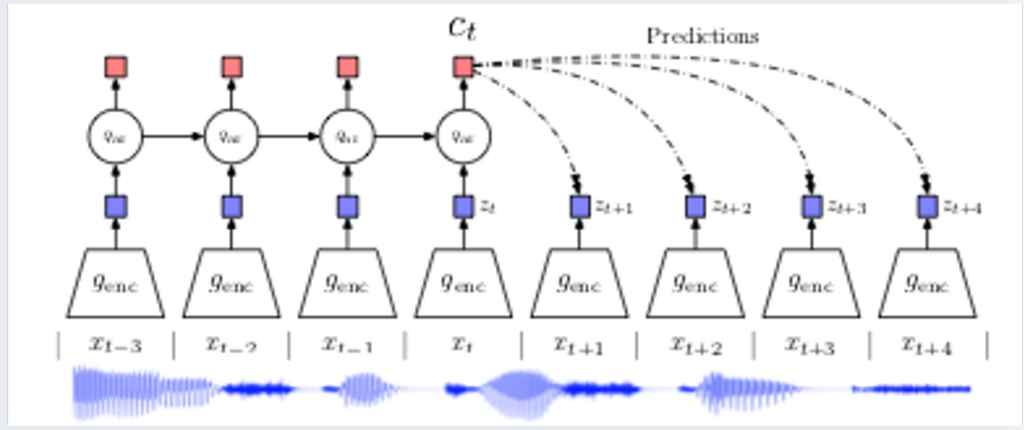
DPGMM[1]



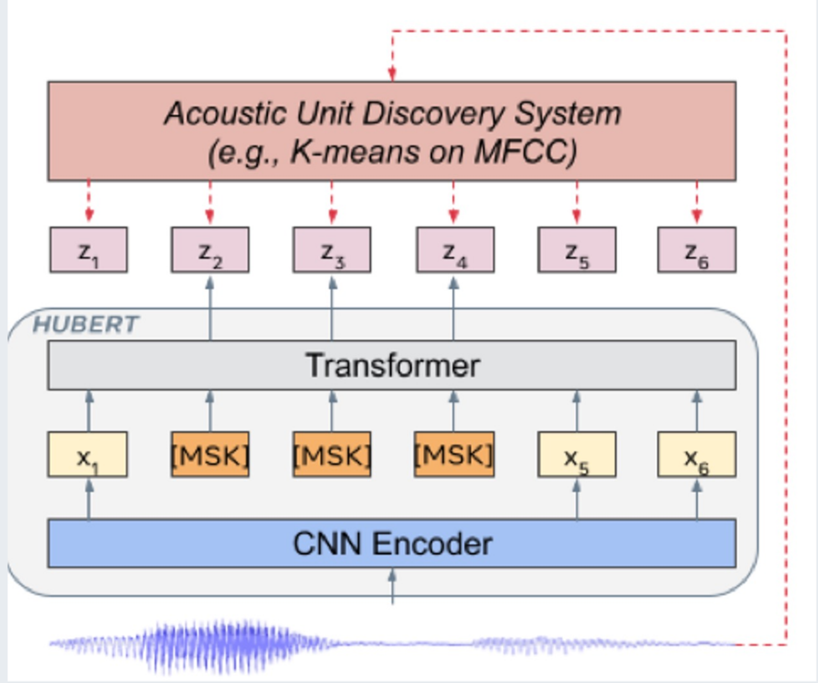
Wavenet autoencoder[2]

Predictive

CPC[3]



HuBERT[4]

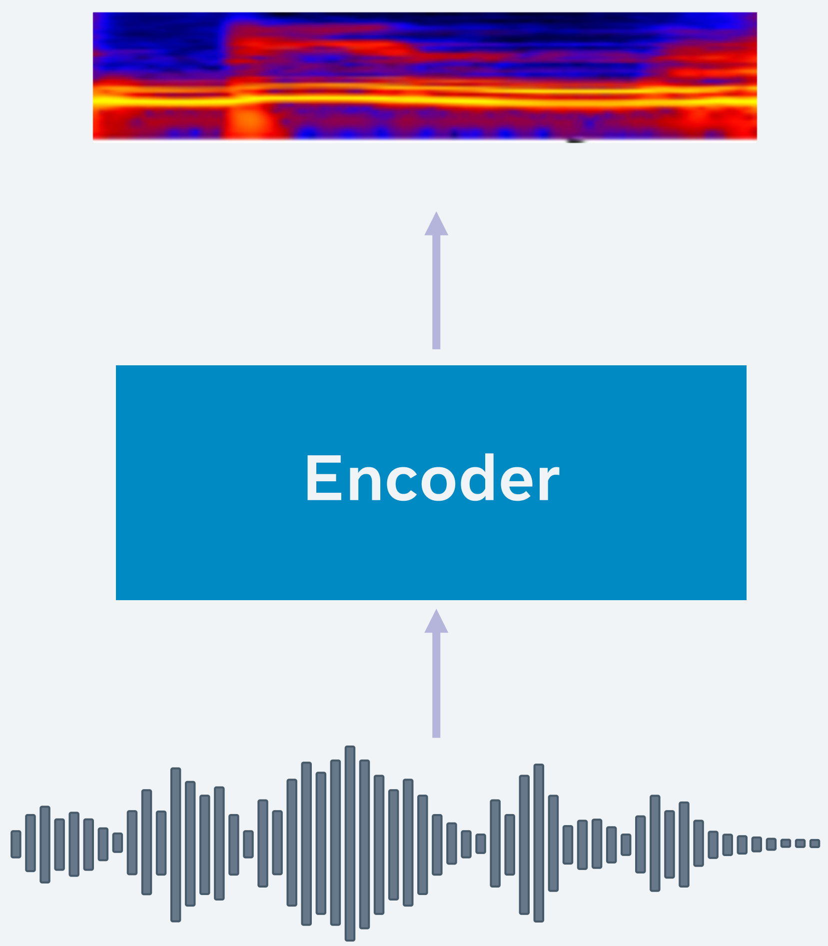


Wav2VEC[5]**, etc**

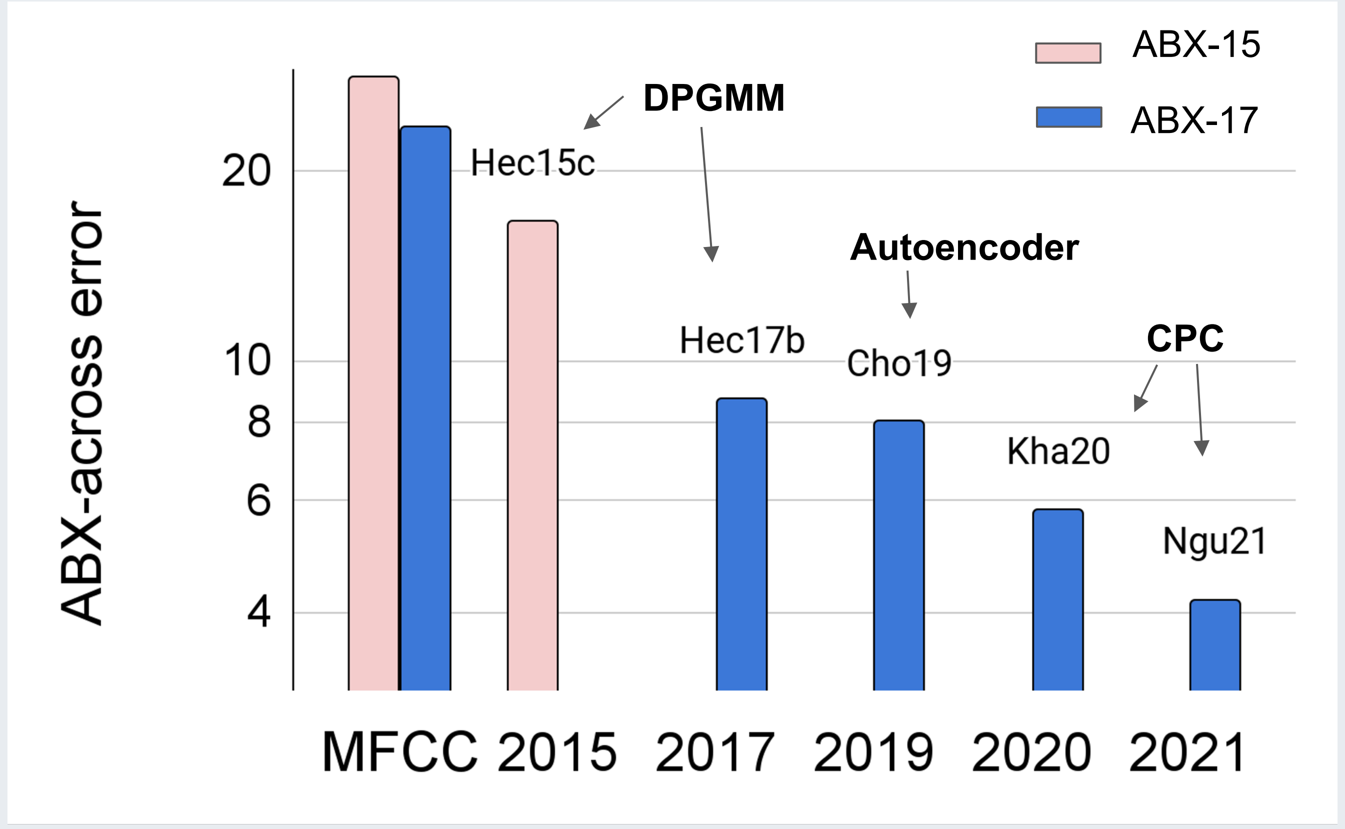
1. Heck et al, 2015, 2017
2. Chorowski et al. 2019
3. Van den Oord, 2018; Kharitonov et al. 2020;
4. Hsu et al, 2021
5. Baevsky et al, 2020

The encoder

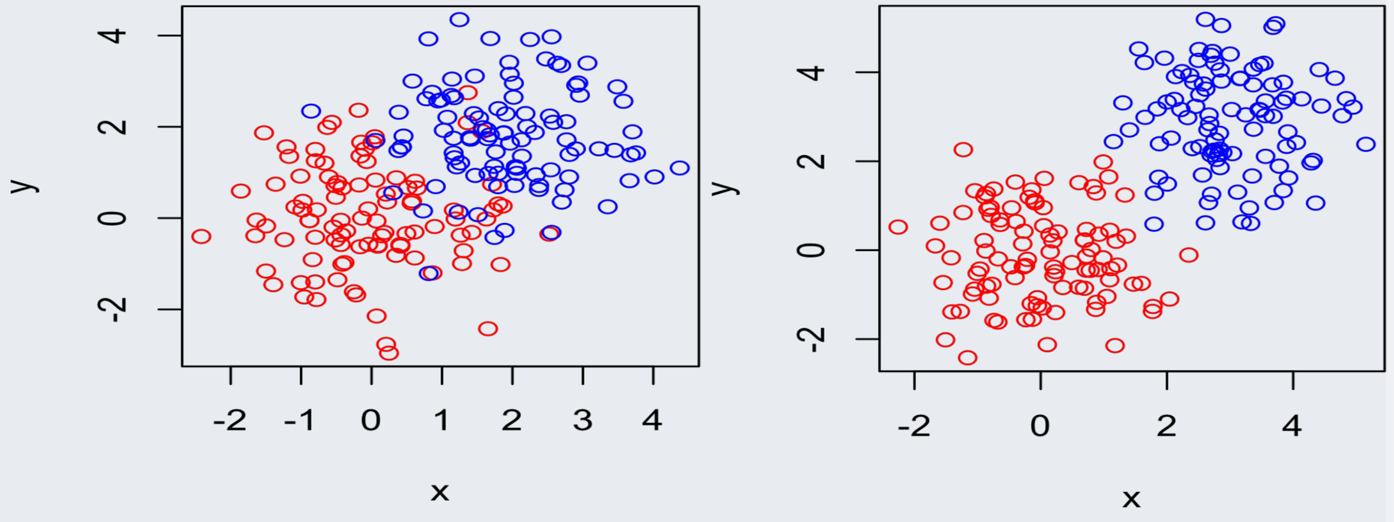
Audio Representation Learning



Leaderboard



Dunbar, Hamilakis, Dupoux (submitted)



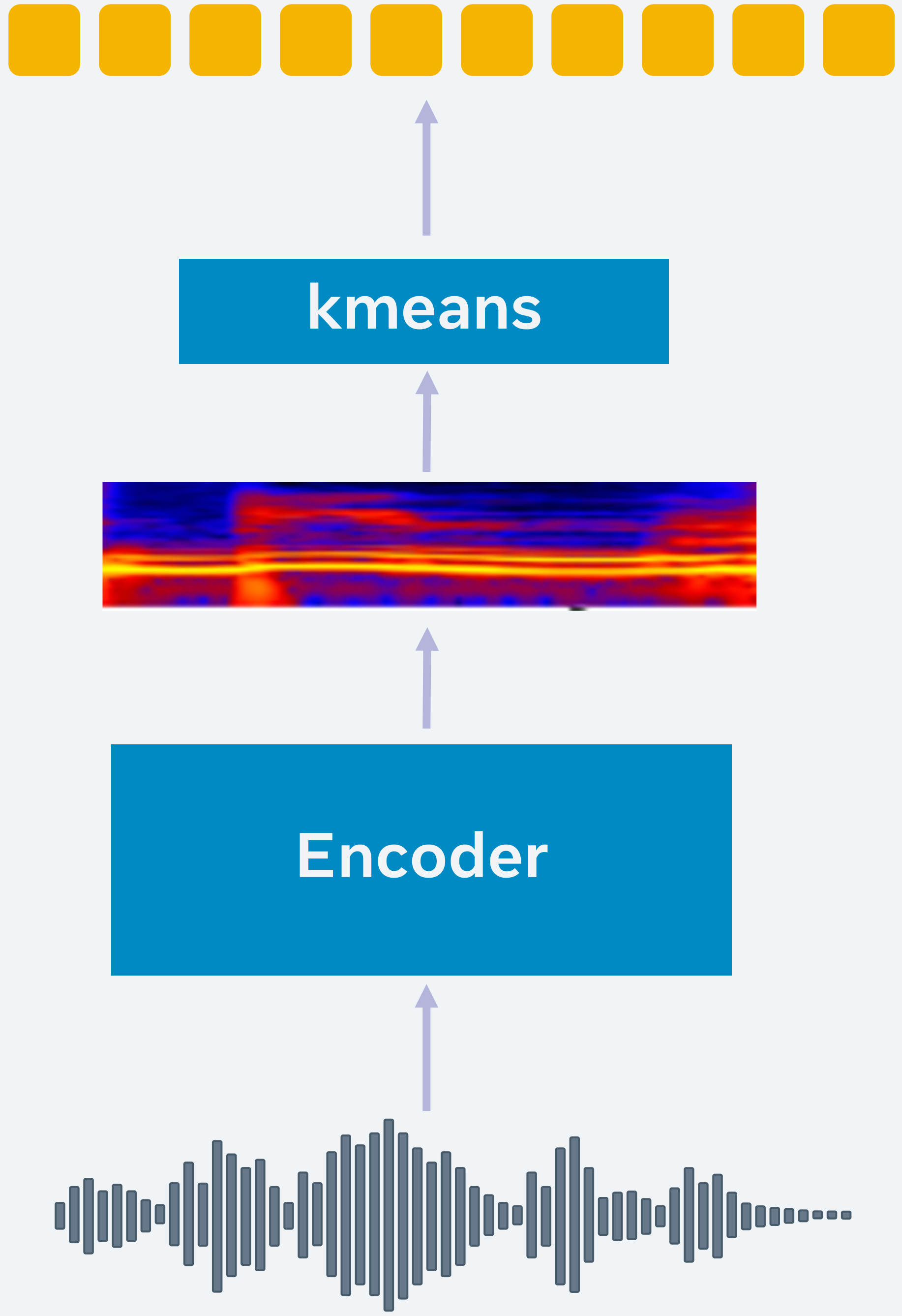
ABX=20%
(2σ)

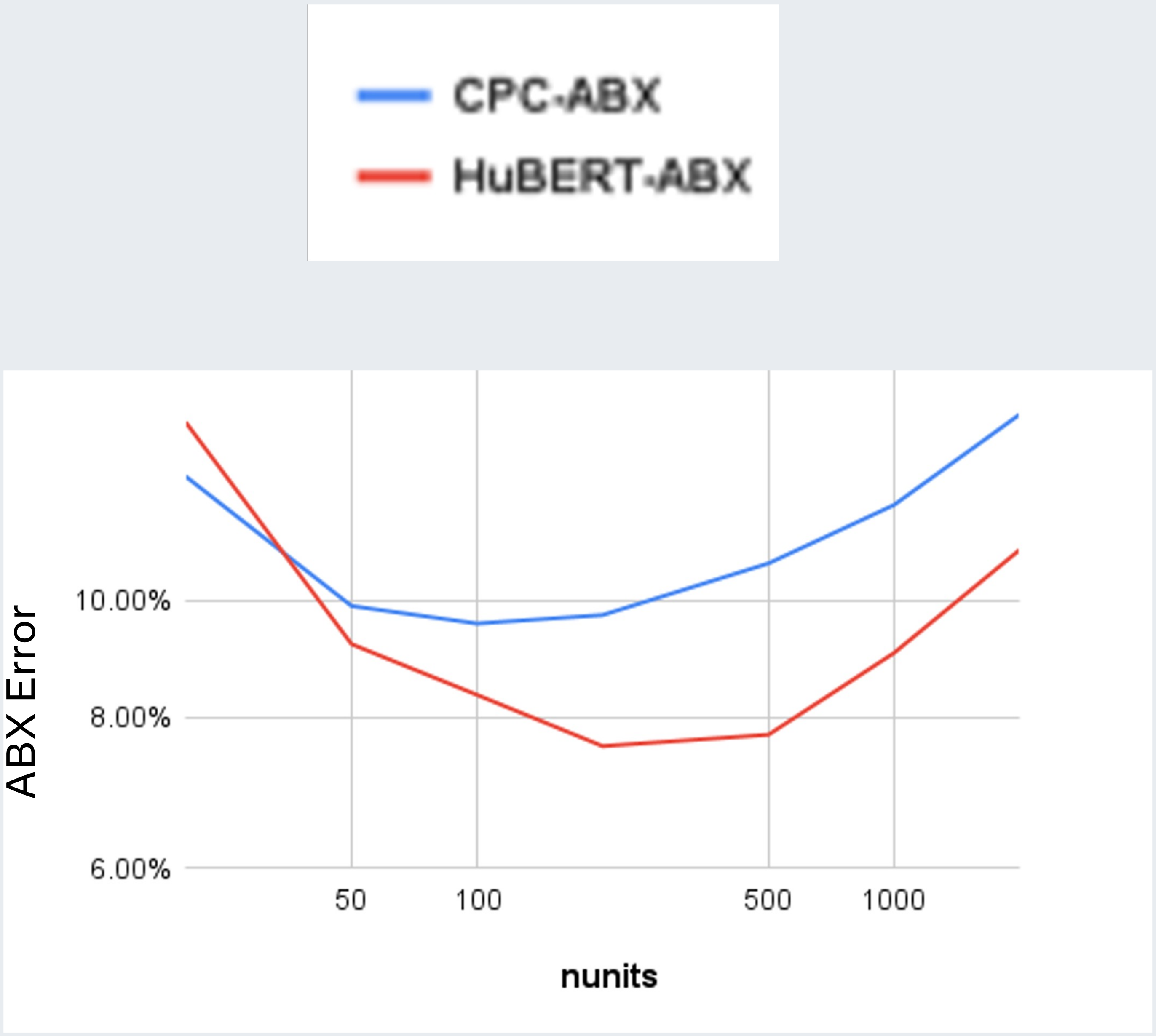
ABX=5%
(2.4σ)

Dunba, Hamilakis, Dupoux (2022)

The encoder

Acoustic Unit Discovery
(discrete representation learning)

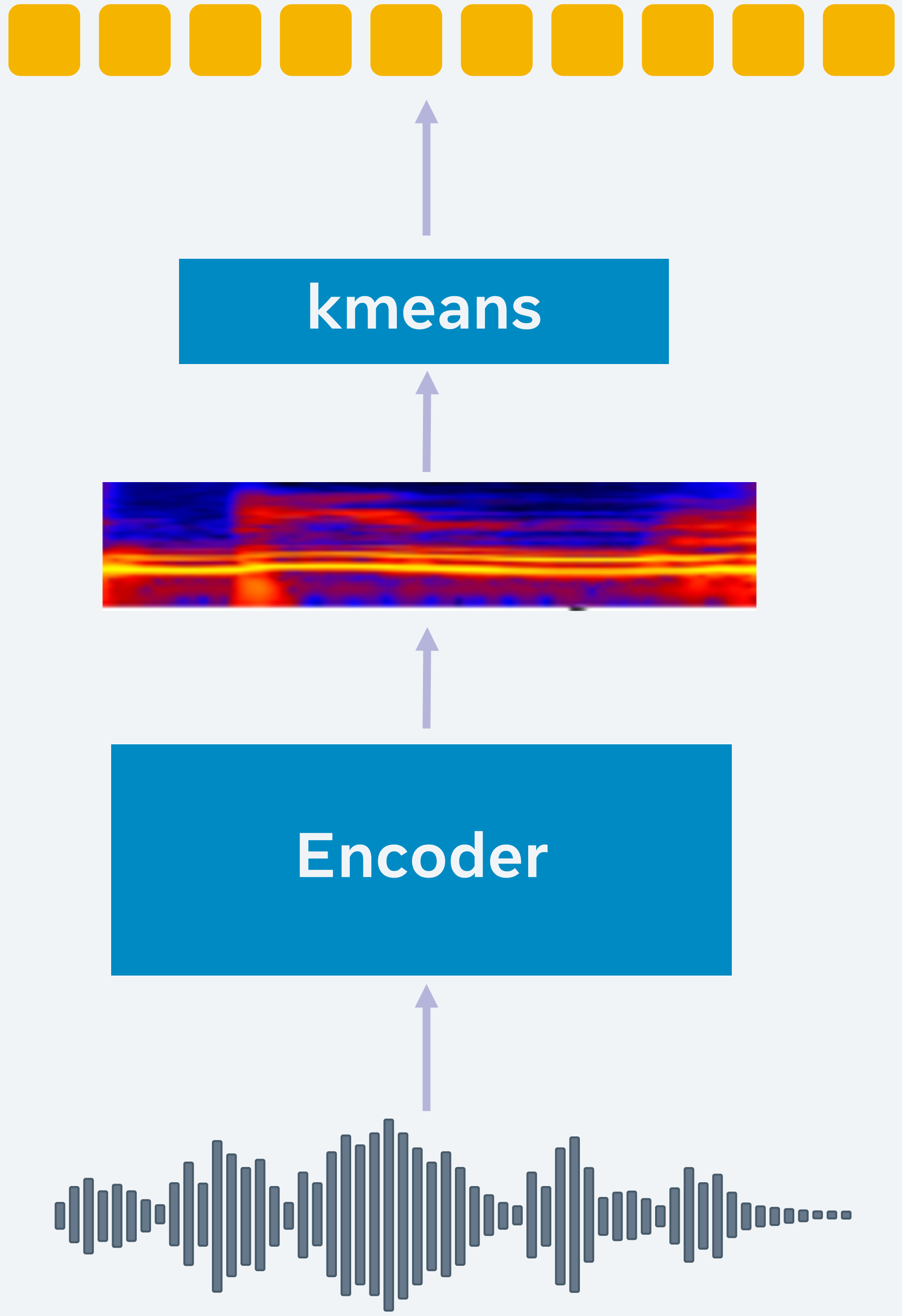




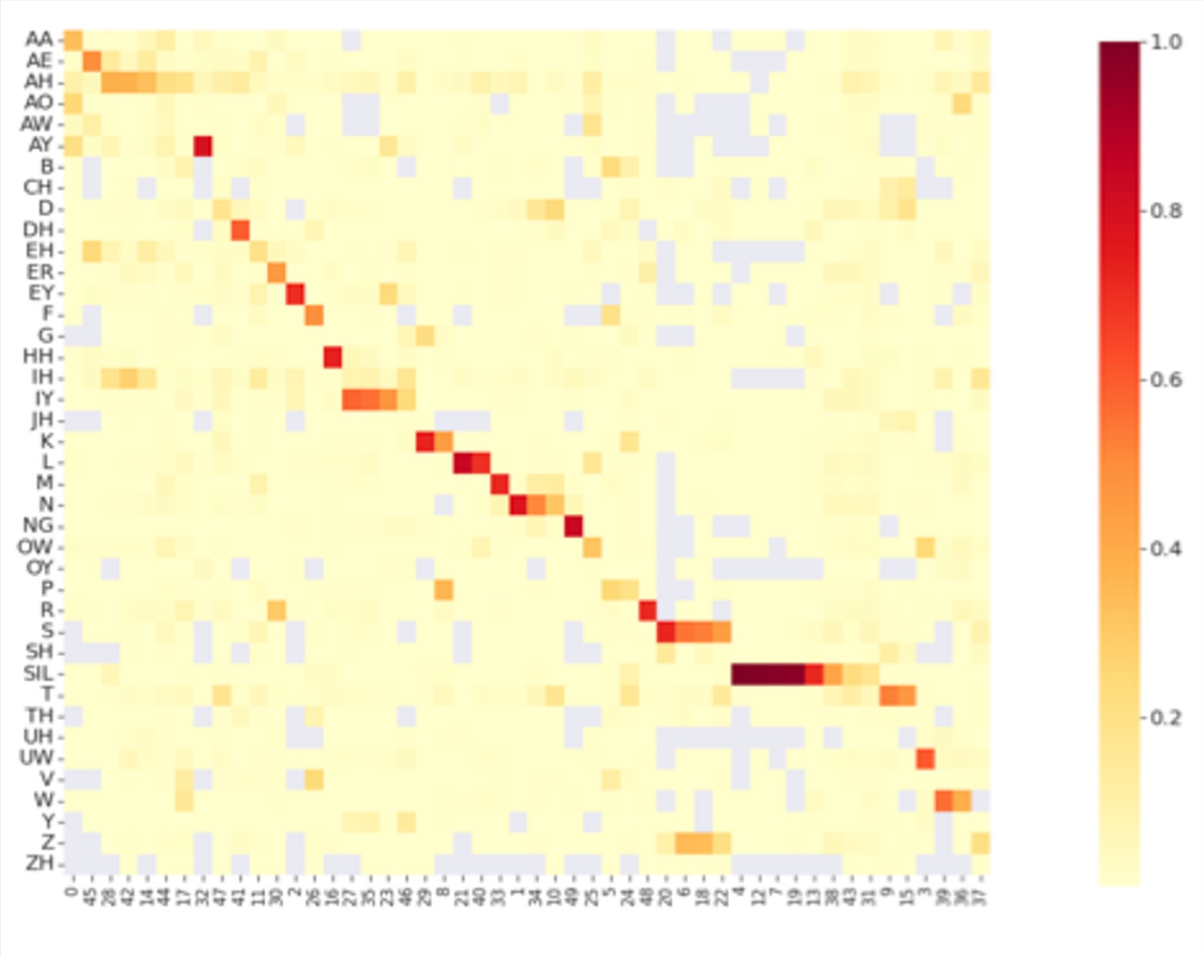
Nguyen et al (2022)

The encoder

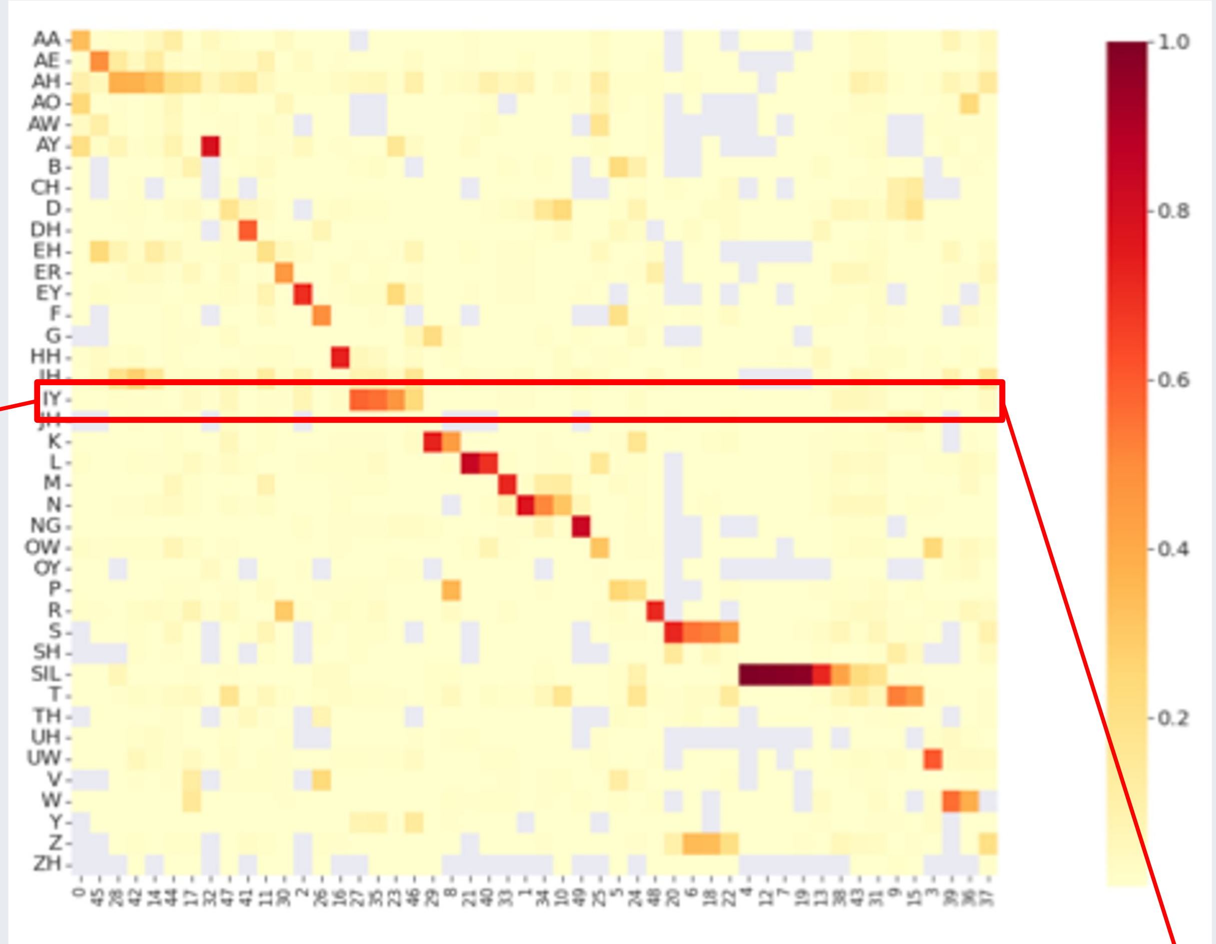
Acoustic Unit Discovery
(discrete representation learning)

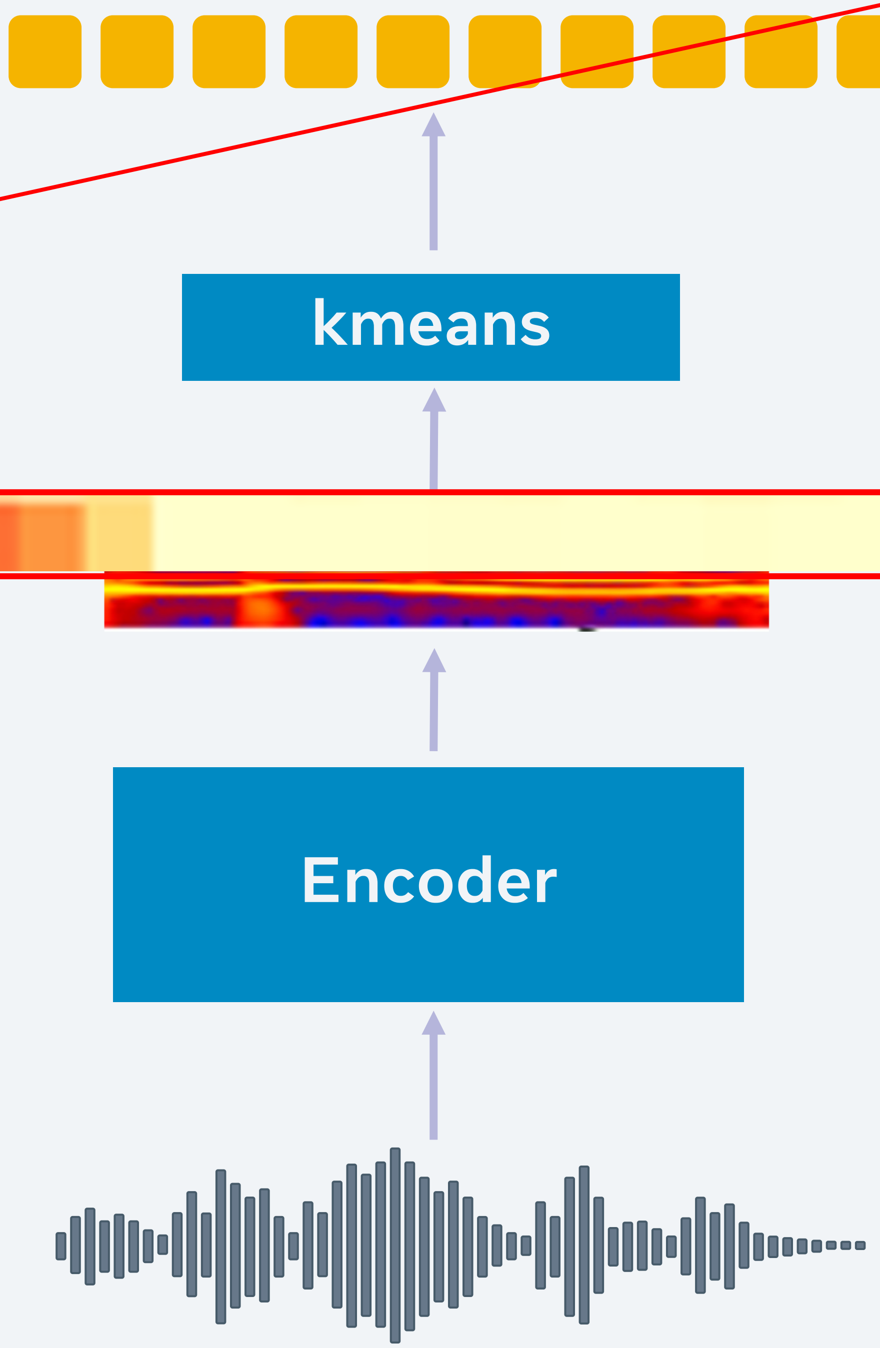


k=50



Nguyen et al (2022)



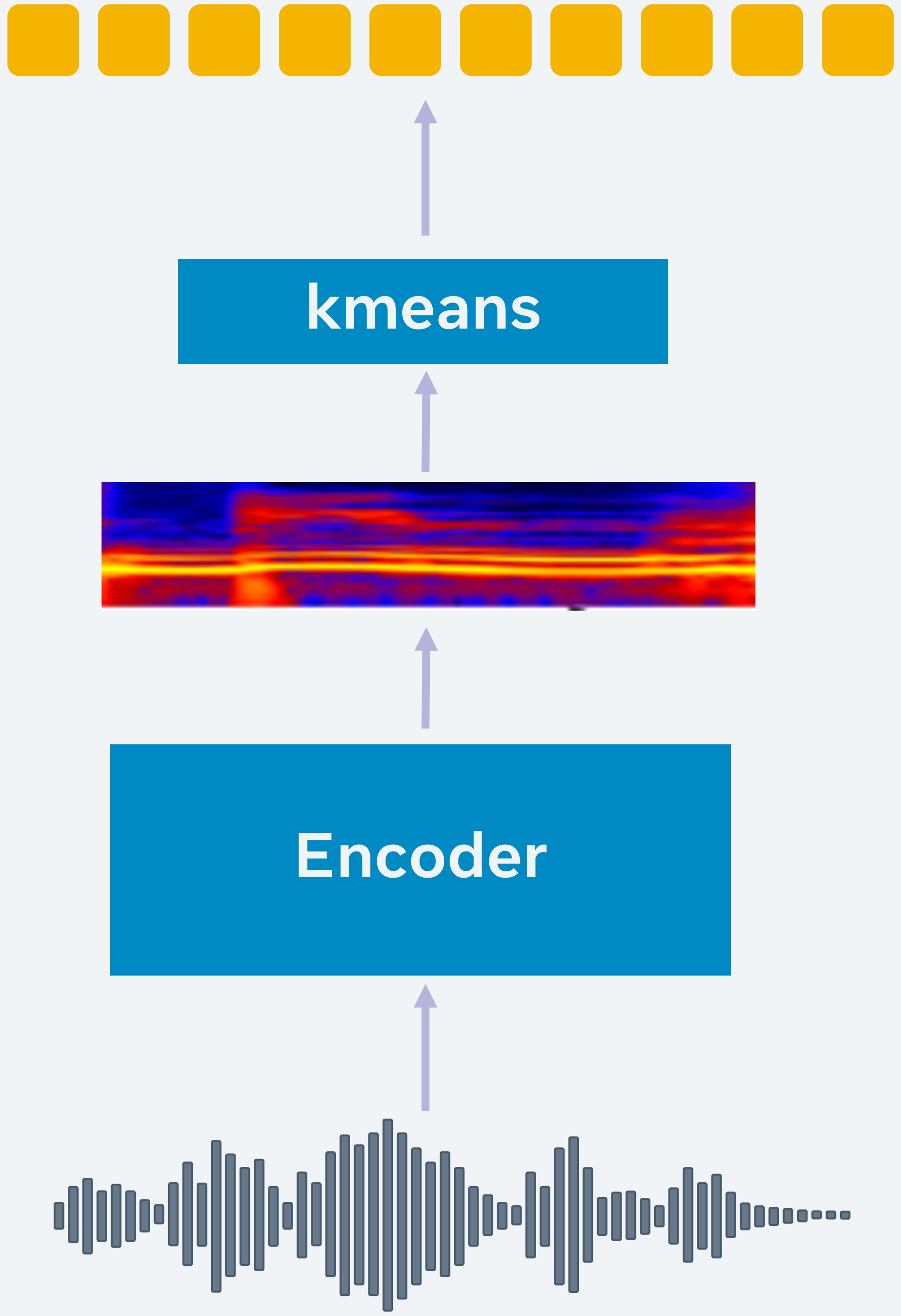


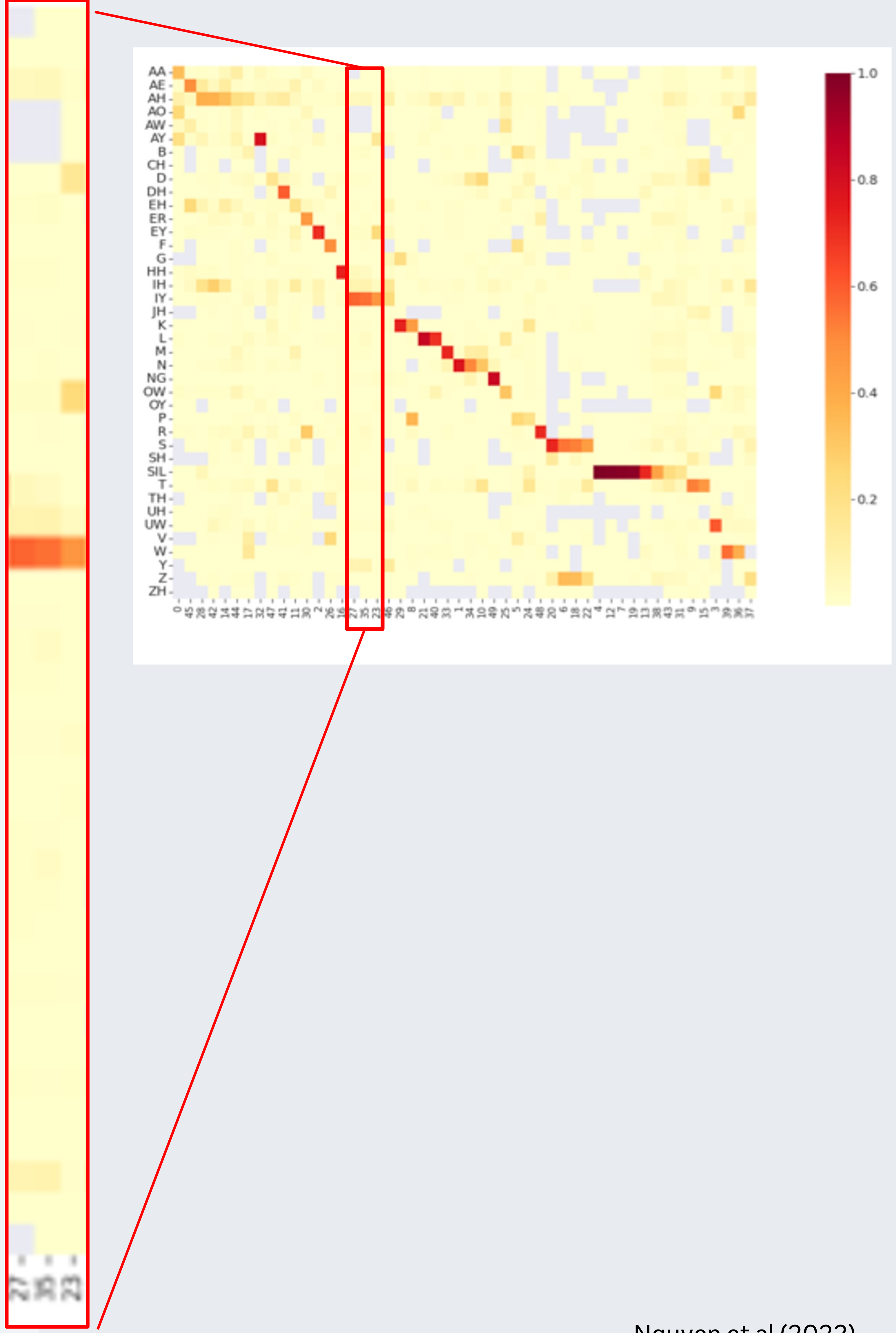
Acoustic Unit Discovery
(discrete representation learning)

Nguyen et al (2022)

The encoder

Acoustic Unit Discovery
(discrete representation learning)

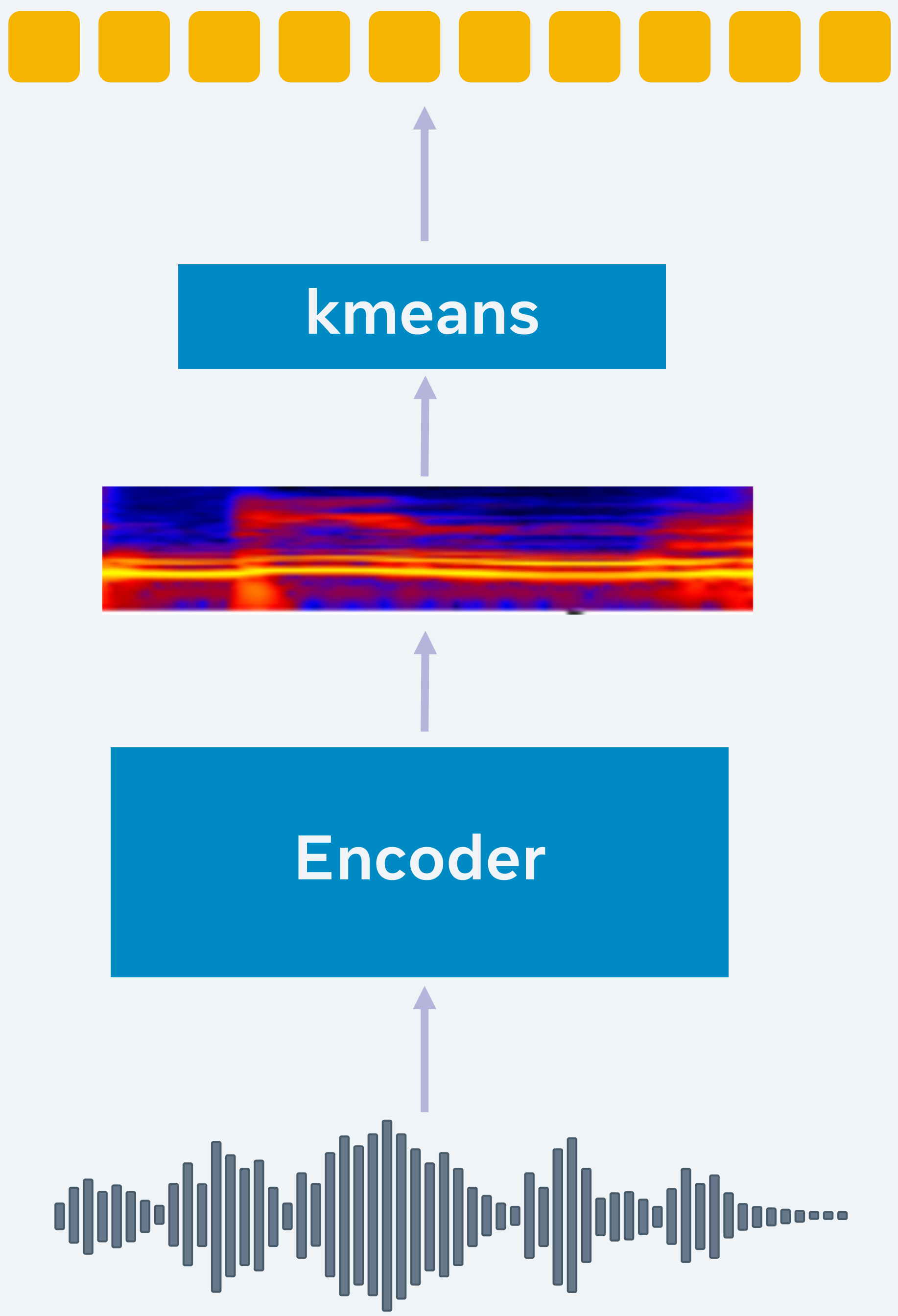


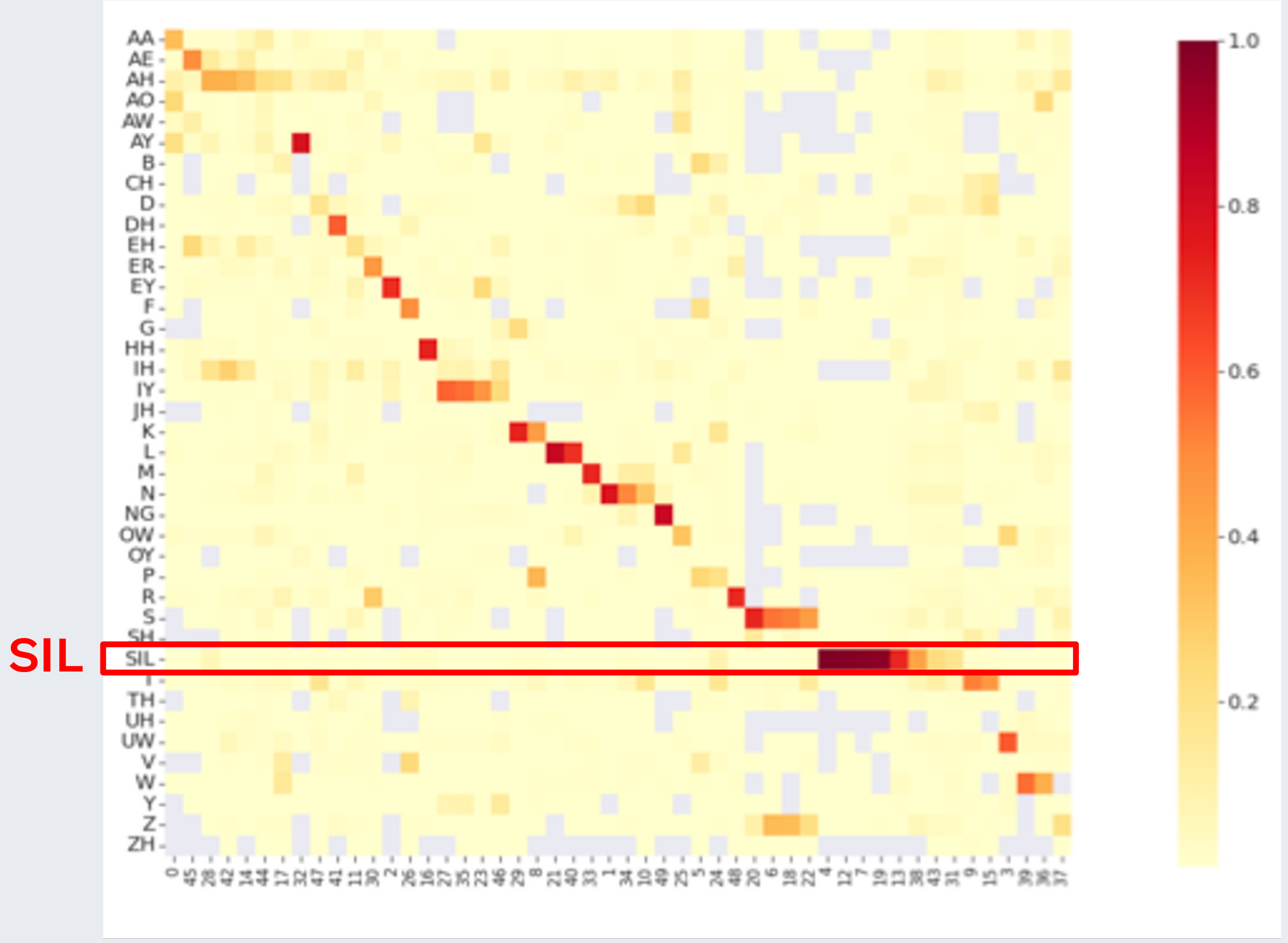


Nguyen et al (2022)

The encoder

Acoustic Unit Discovery
(discrete representation learning)

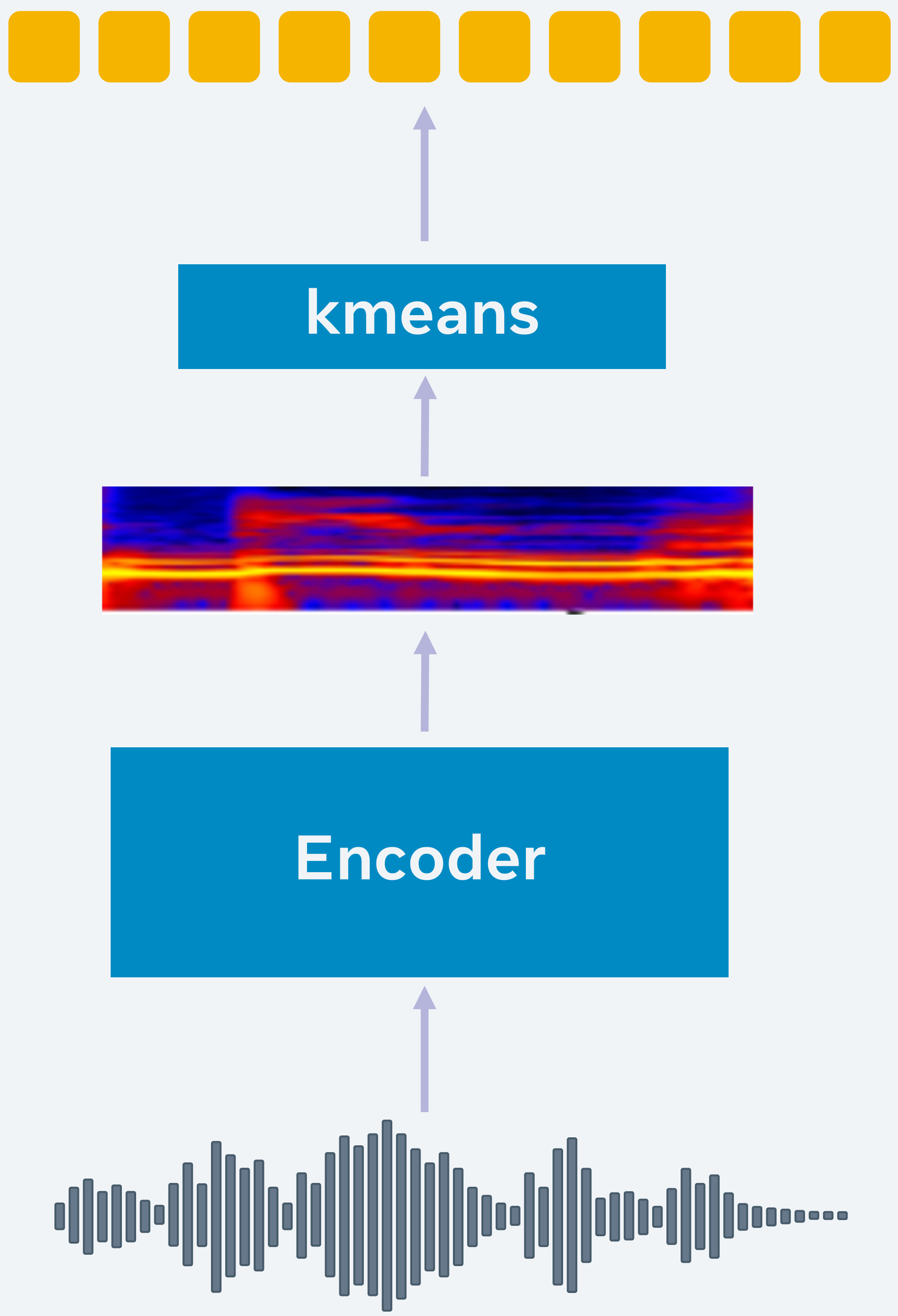


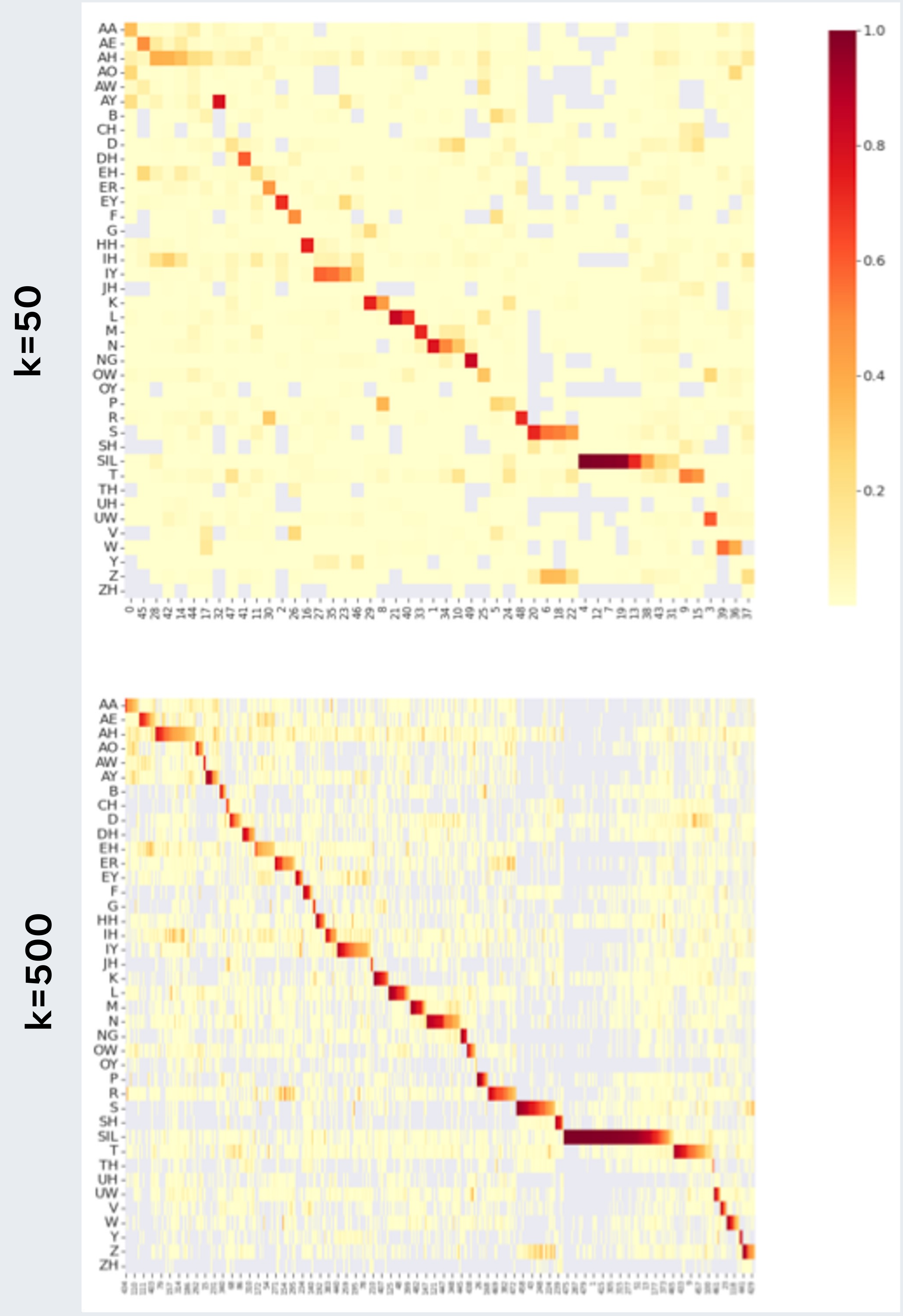


Nguyen et al (2022)

The encoder

Acoustic Unit Discovery
(discrete representation learning)

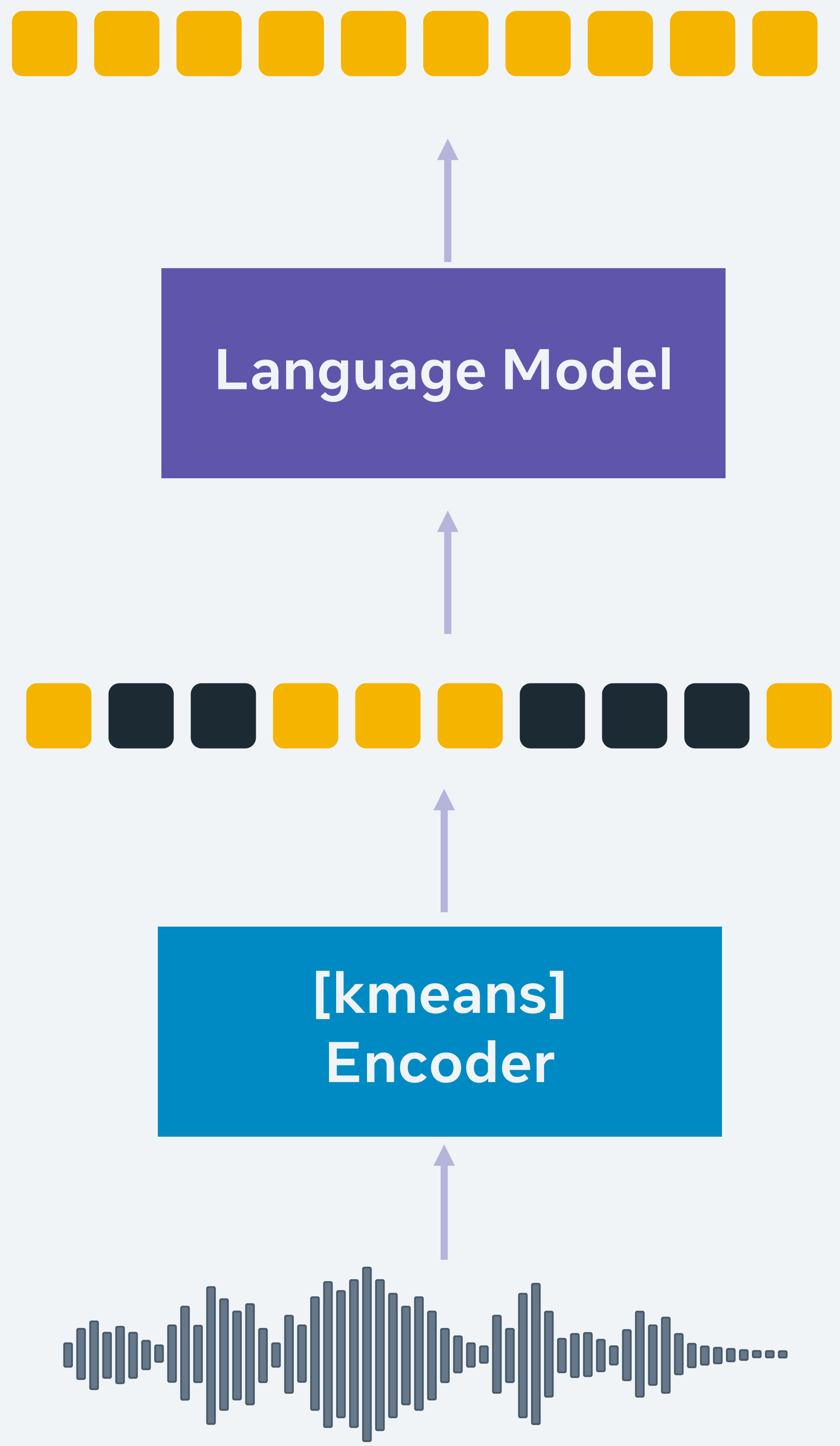




Nguyen et al (2022)

The language model

Spoken Language Modeling



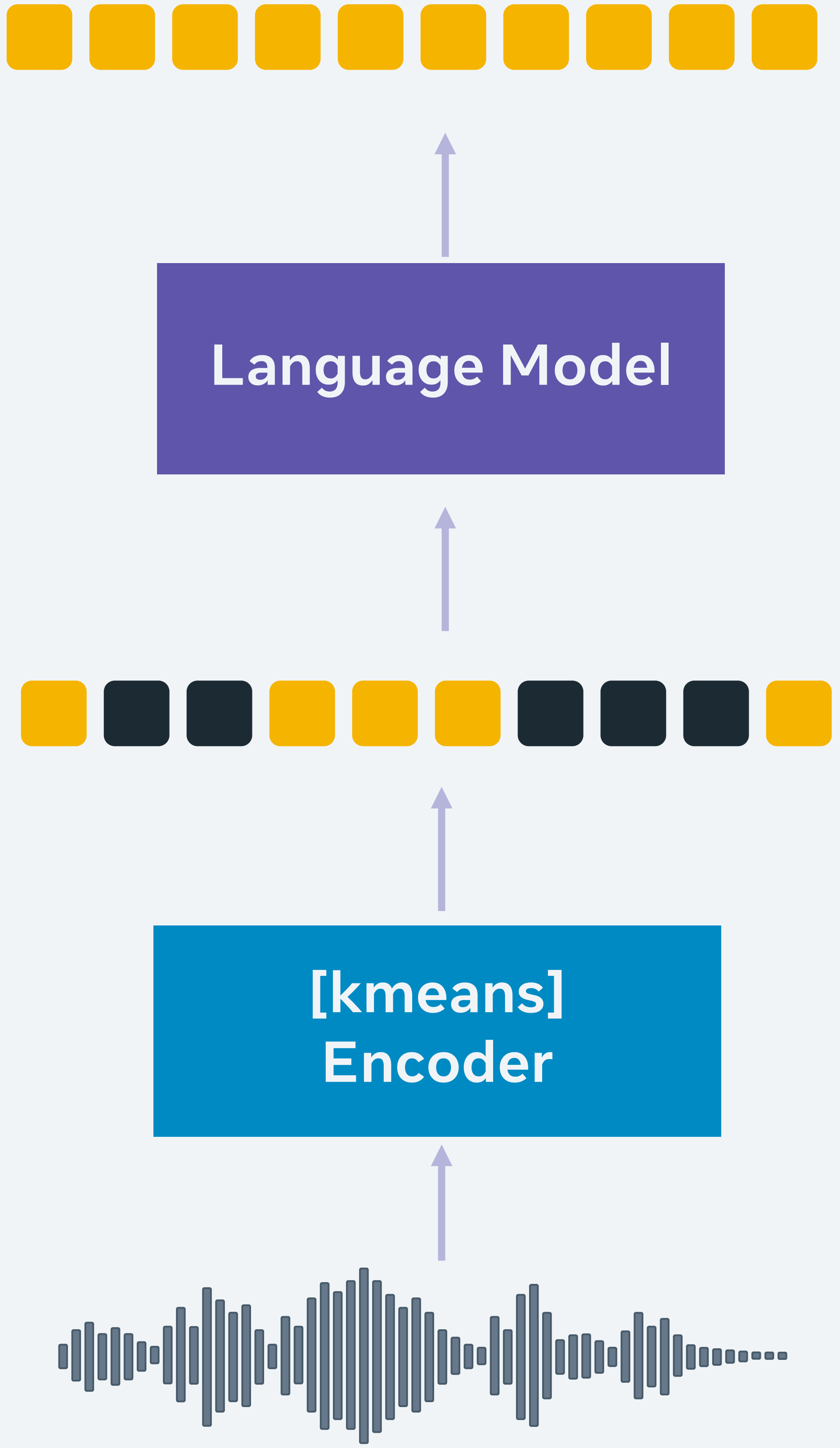
ZRC Task 4
Learn the probabilistic distribution of speech

Evaluation:

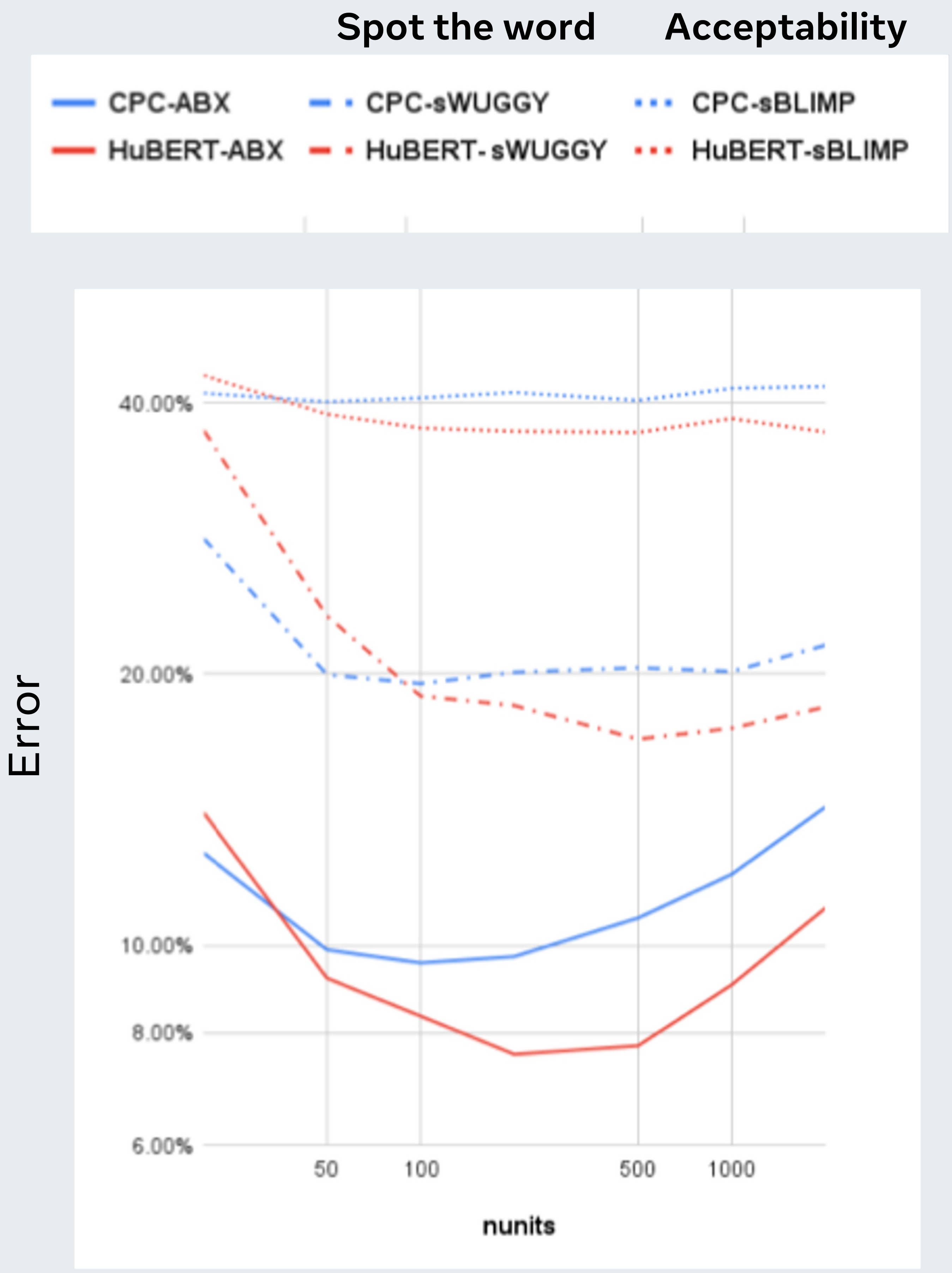
Levels	Tasks
Syntactic	**accept . judgment** *"they like" vs "they likes"*
Lexical	**spot-the-word** *"blick" vs "brick"*

The language model

Spoken Language Modeling



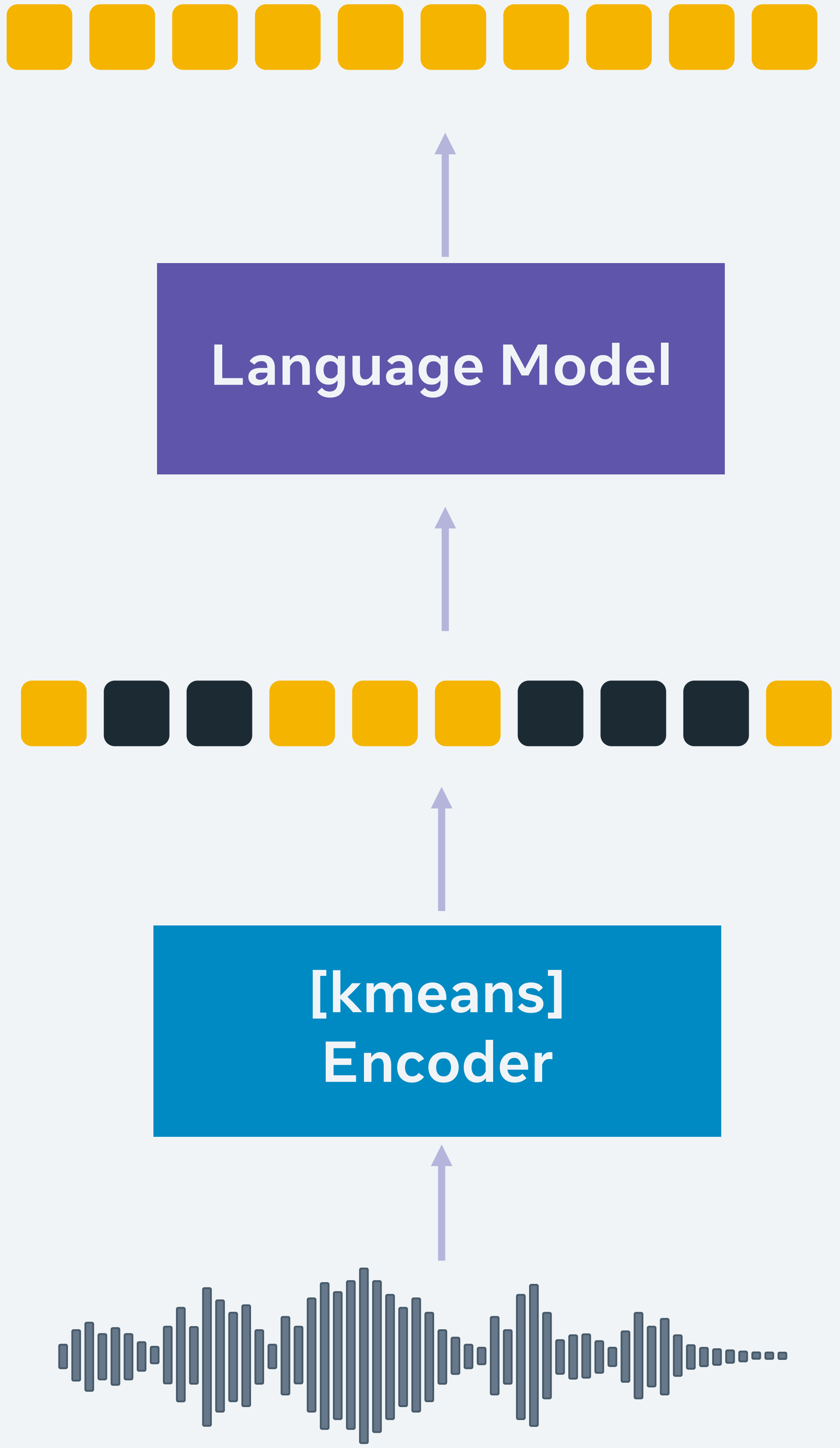
Leaderboard



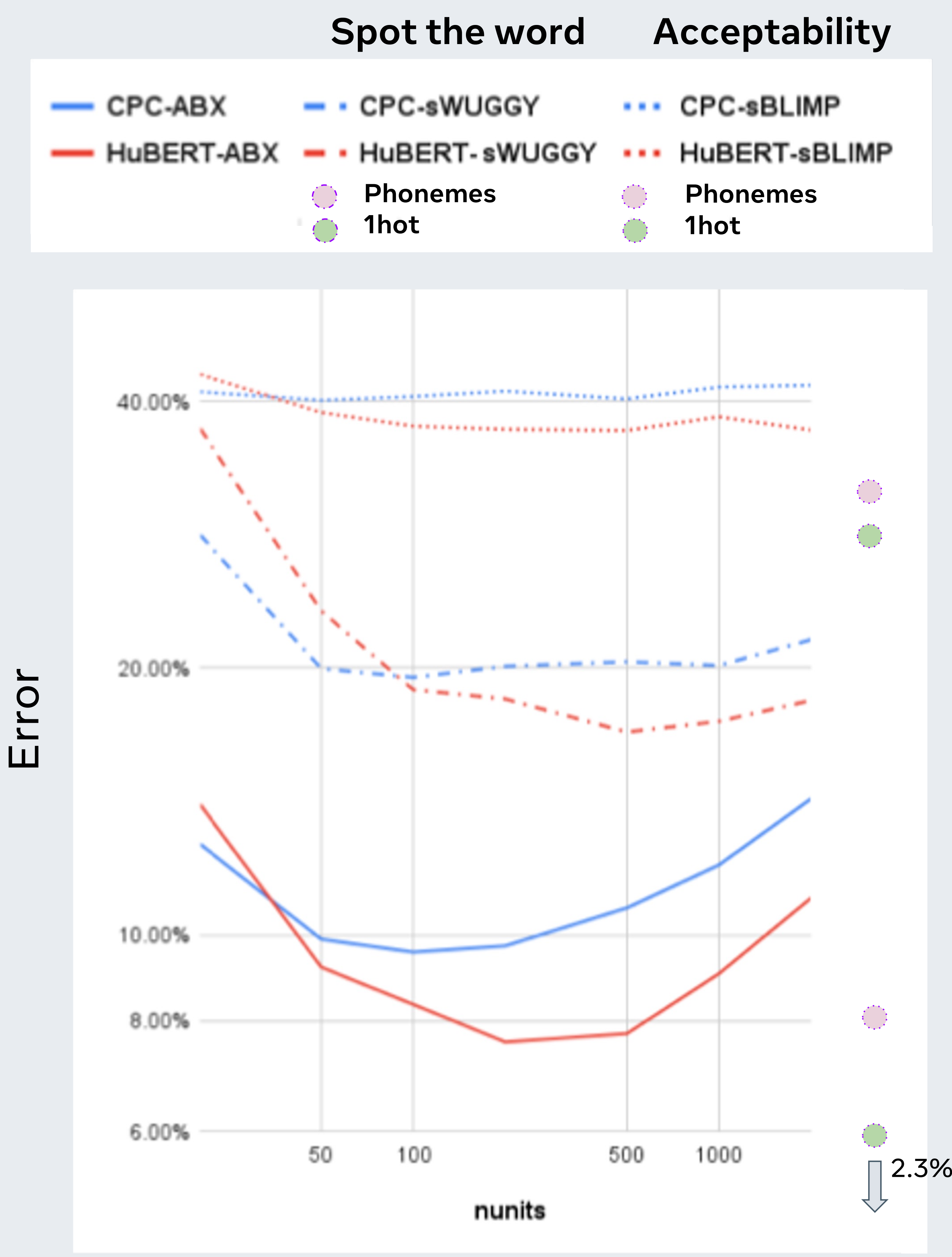
Nguyen et al (2022)

The language model

Spoken Language Modeling



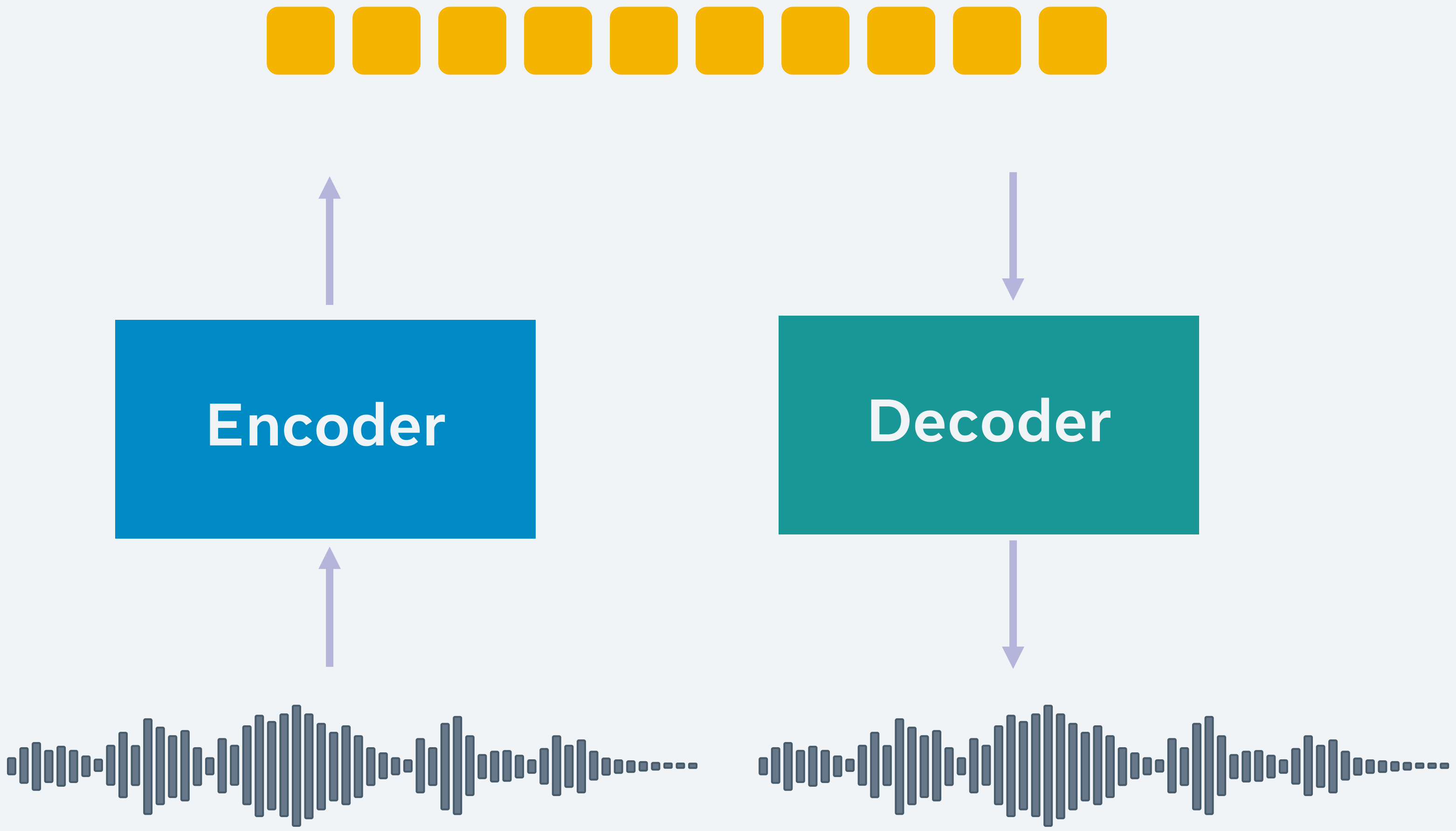
Leaderboard



Nguyen et al (2022)

The decoder

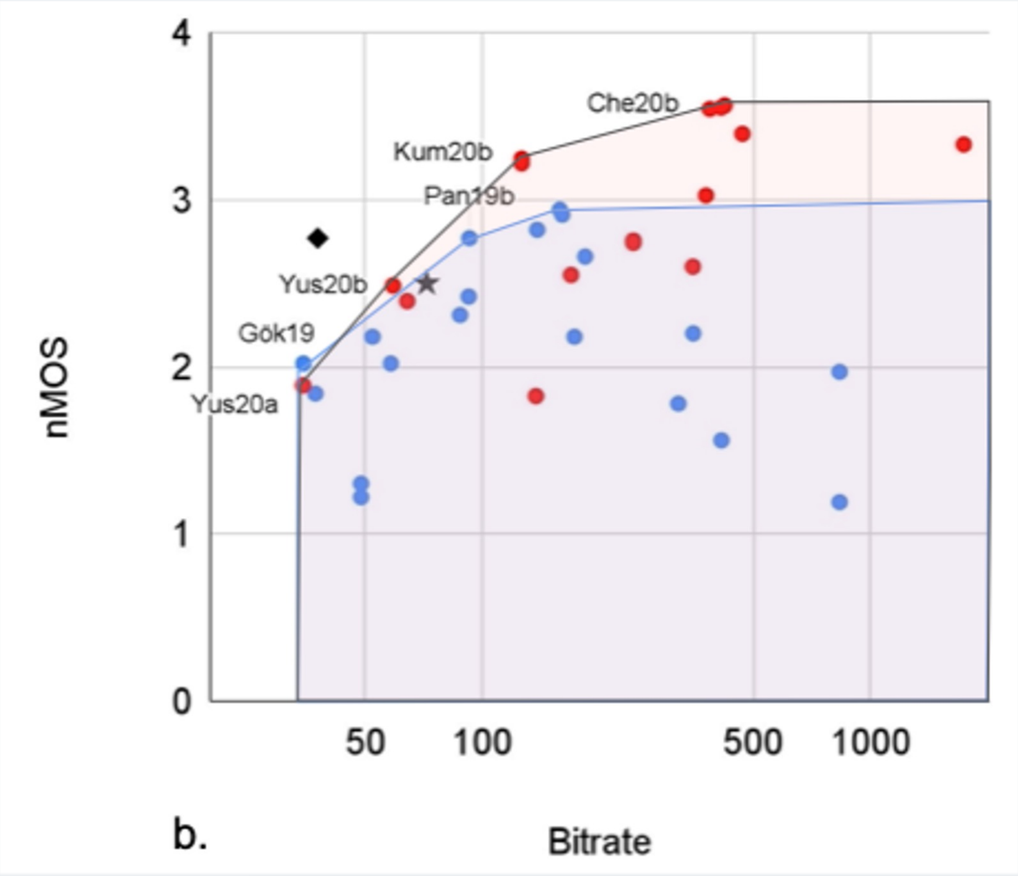
Discrete resynthesis



ZRC Task 3
Resynthetize speech from a discrete code

Evaluation:

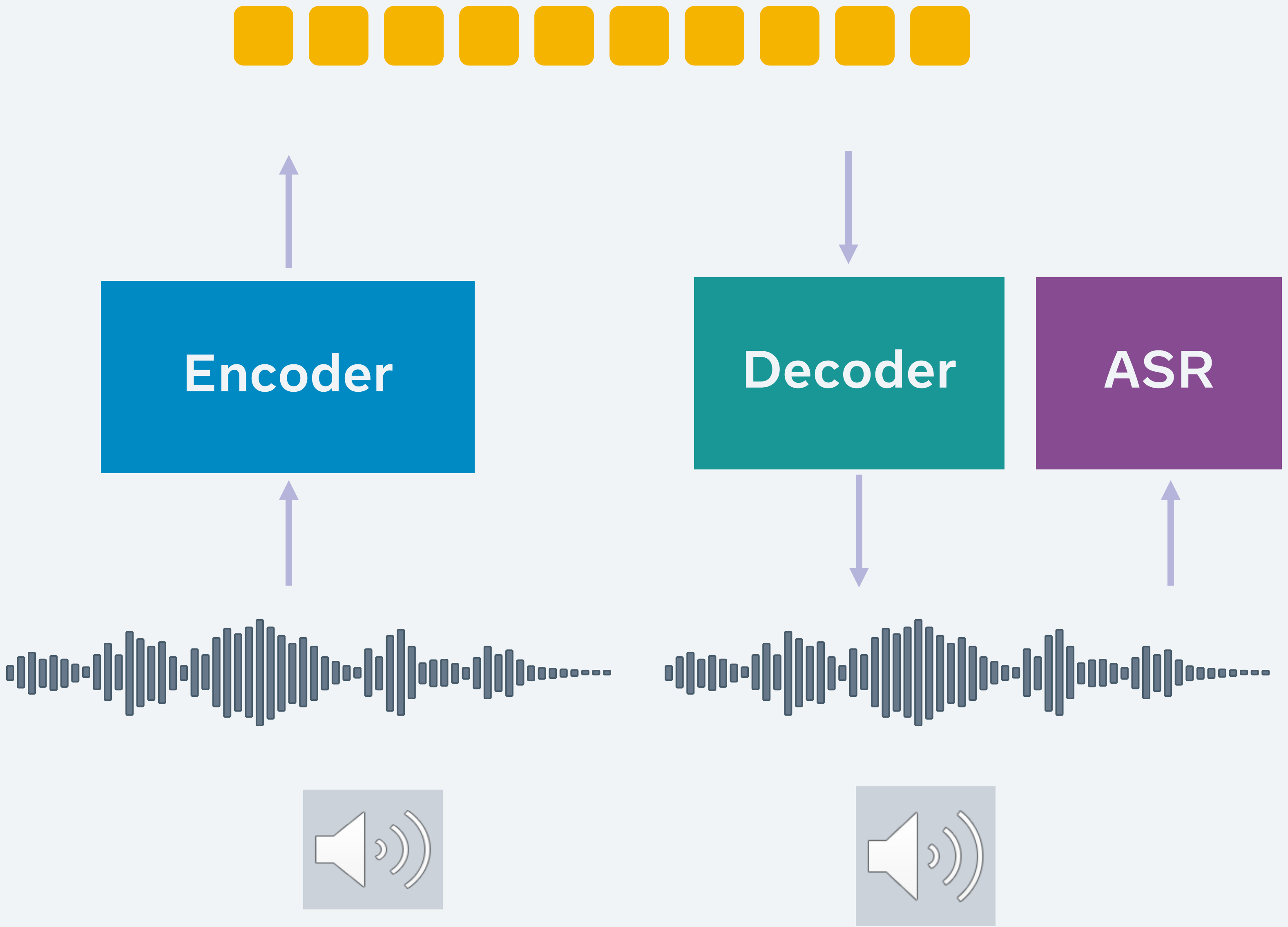
- Human (MOS)



Dunbar, Hamilakis, Dupoux (2022)

The decoder

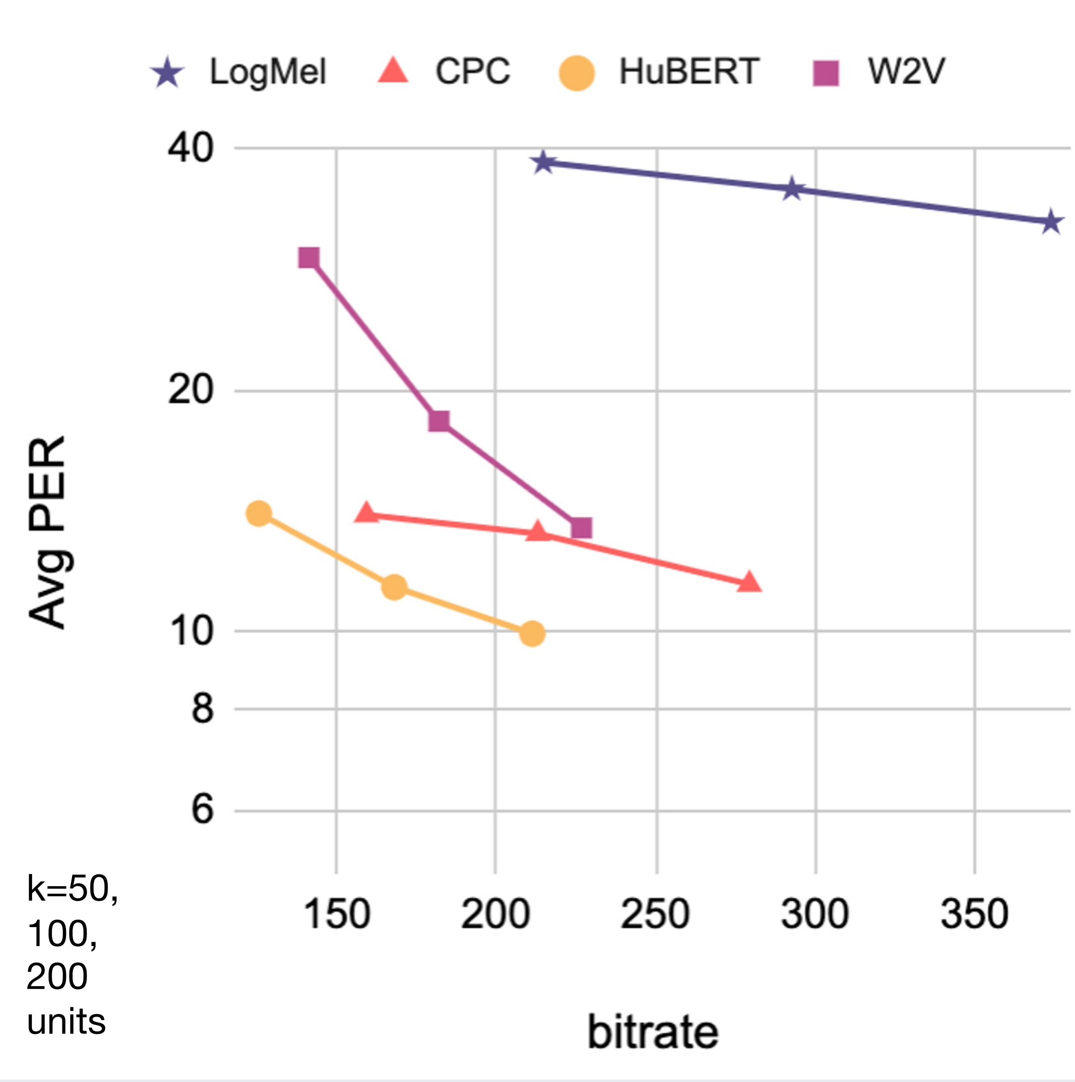
Discrete resynthesis



ZRC Task 3
Resynthetize speech from a discrete code

Evaluation:

- PER

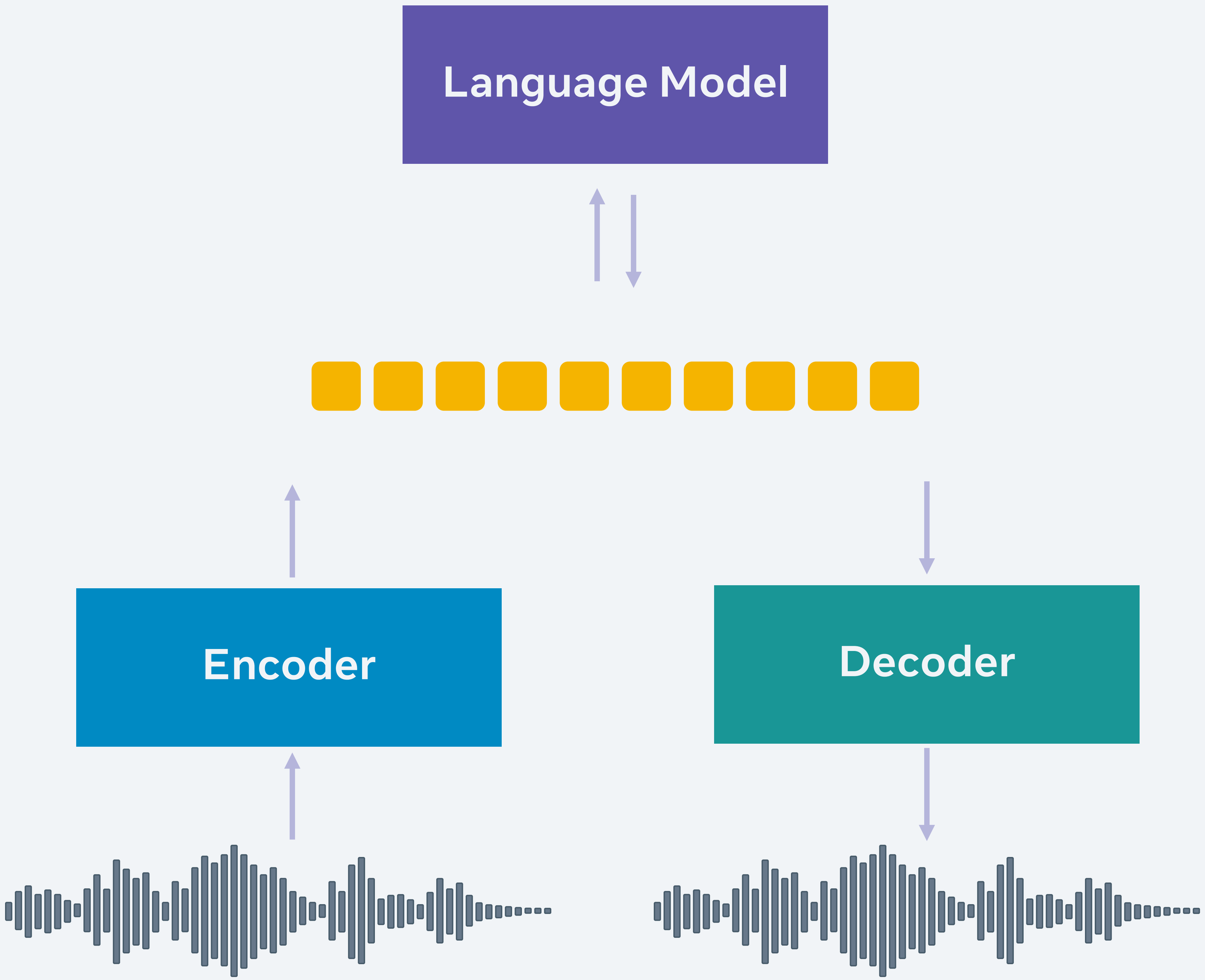


Correlation between PER and MOS, R=.90-.95

Lakhotia et al (2021). Generative spoken language modeling. TACL.

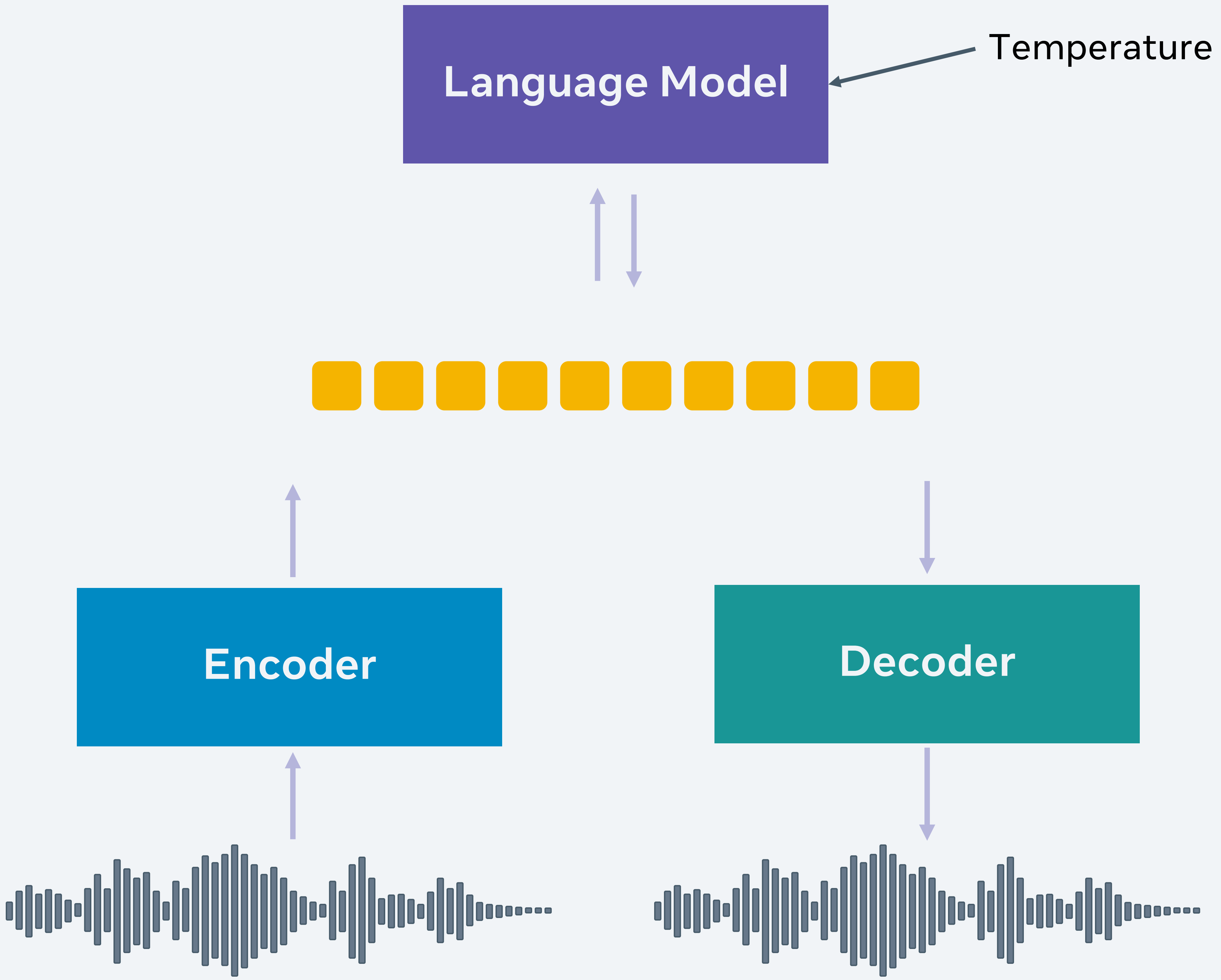
Generative Spoken Language Modeling

Putting all together



Generative Spoken Language Modeling

Putting all together



Unconditional generation (low temperature)

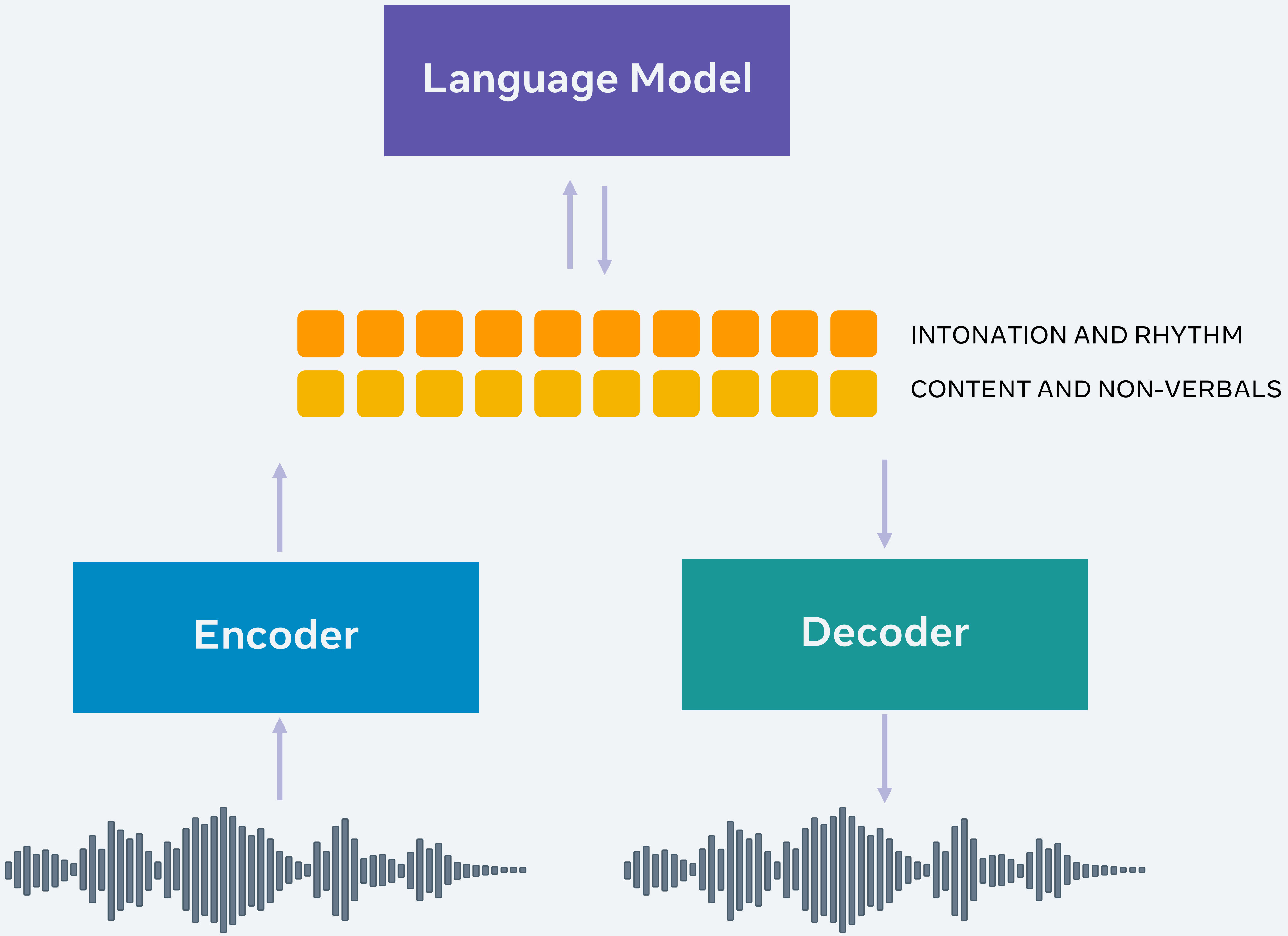
Unconditional generation (high temperature)

Unconditional generation (medium temperature)

Lakhotia et al (2021). Generative spoken language modeling. TACL.

Prosodic Generative Spoken Language Modeling

Expressive language modeling



Kharitonov et al (2021). Text-Free Prosody-Aware Generative Spoken Language Modeling; ACL

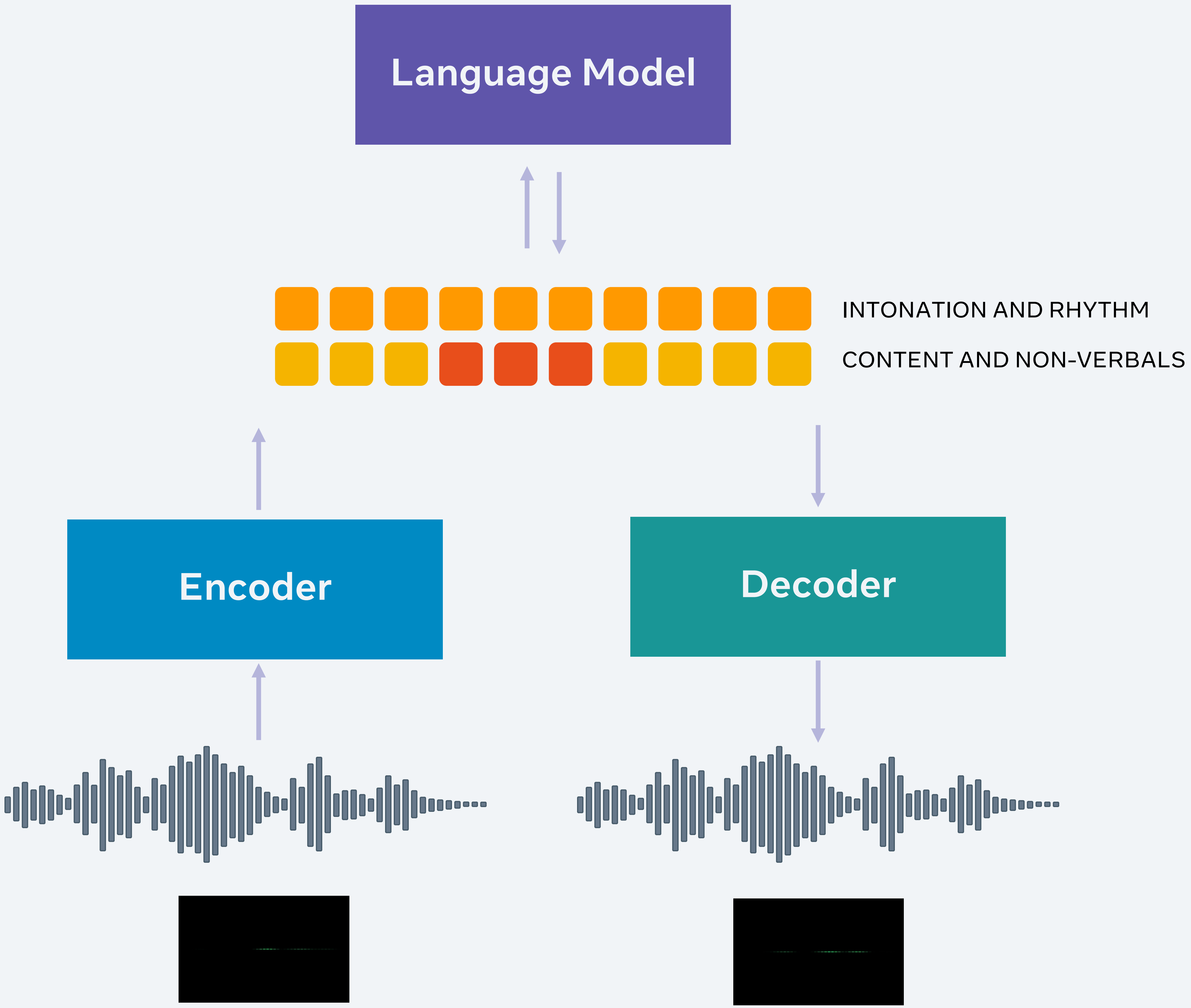
Conditional samples

Kharitonov et al (2021). Text-Free Prosody-Aware Generative Spoken Language Modeling. ACL

More samples: https://speechbot.github.io/pgslm

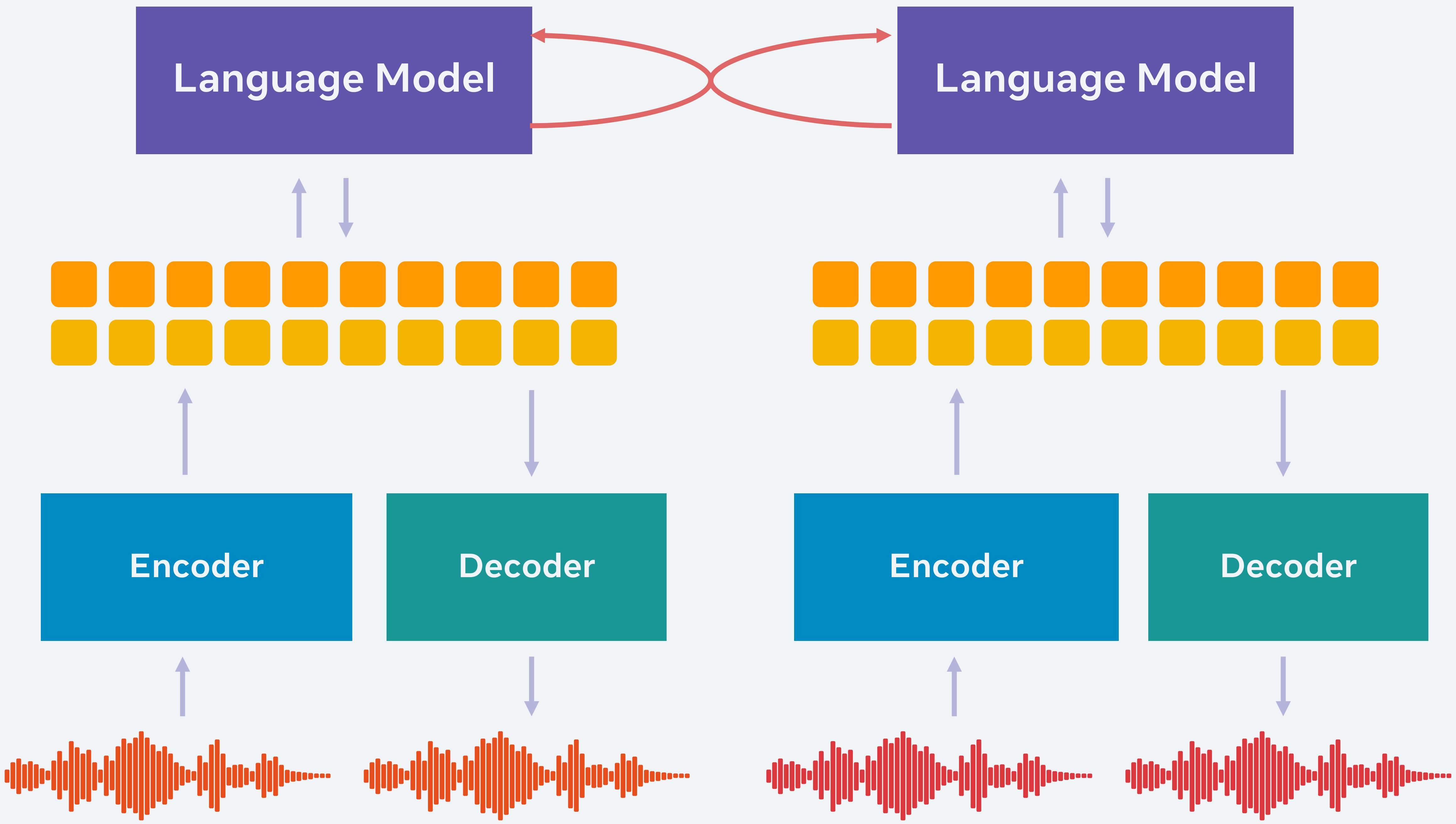
Speech-to-speech applications

Emotion conversion

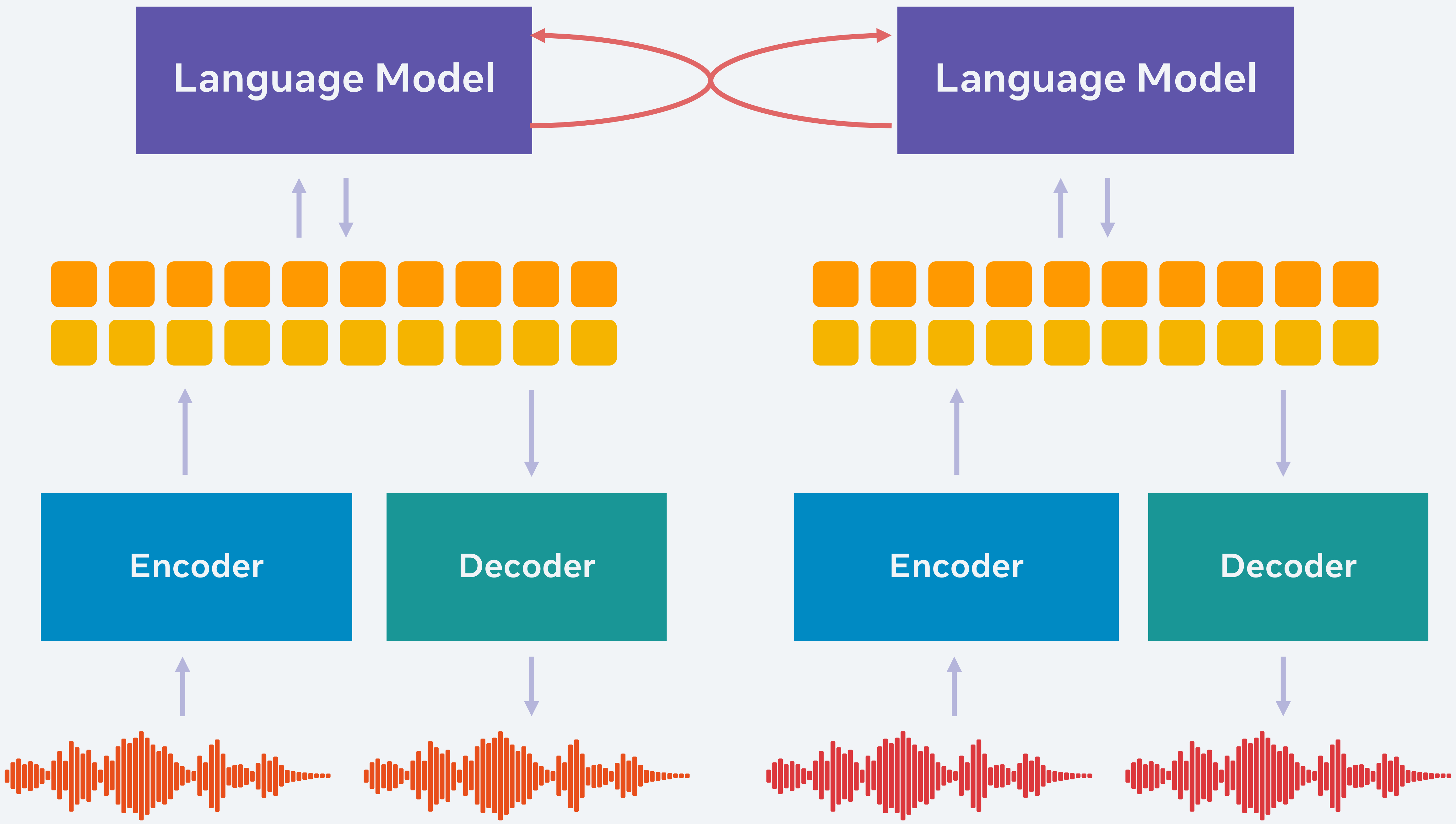


Kreuk et akl (2021).

Dialogue modeling



Dialogue modeling

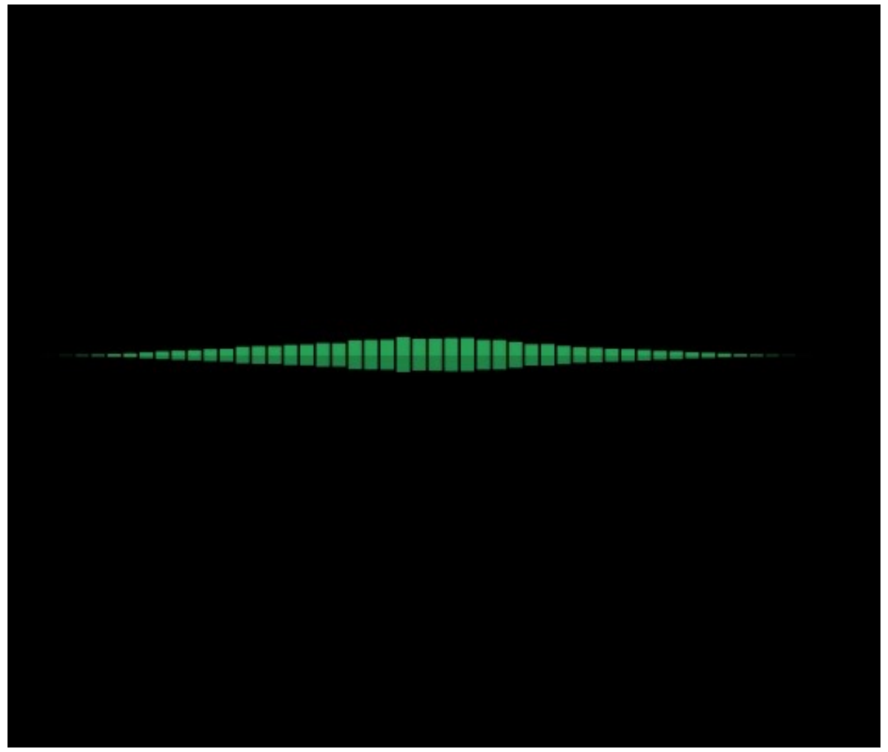


4 second prompt with real humans

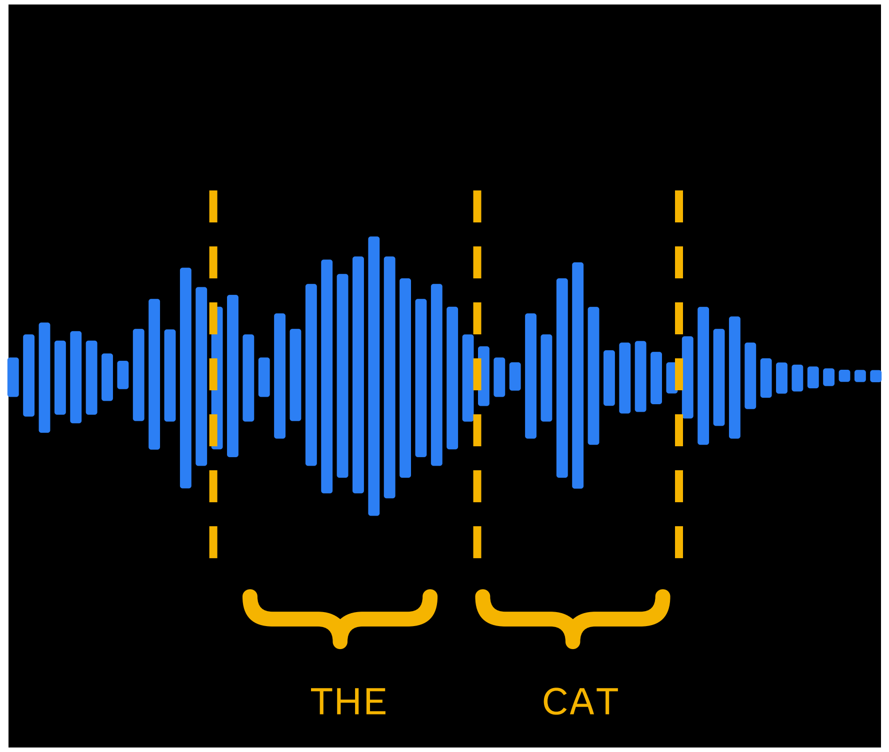
Automatic continuation

Nguyen et al (2022) Generative Spoken Dialogue Modeling

Challenges



Noise and variability of real-world audio[1]



Meaningful segment discovery

Data collection & curation

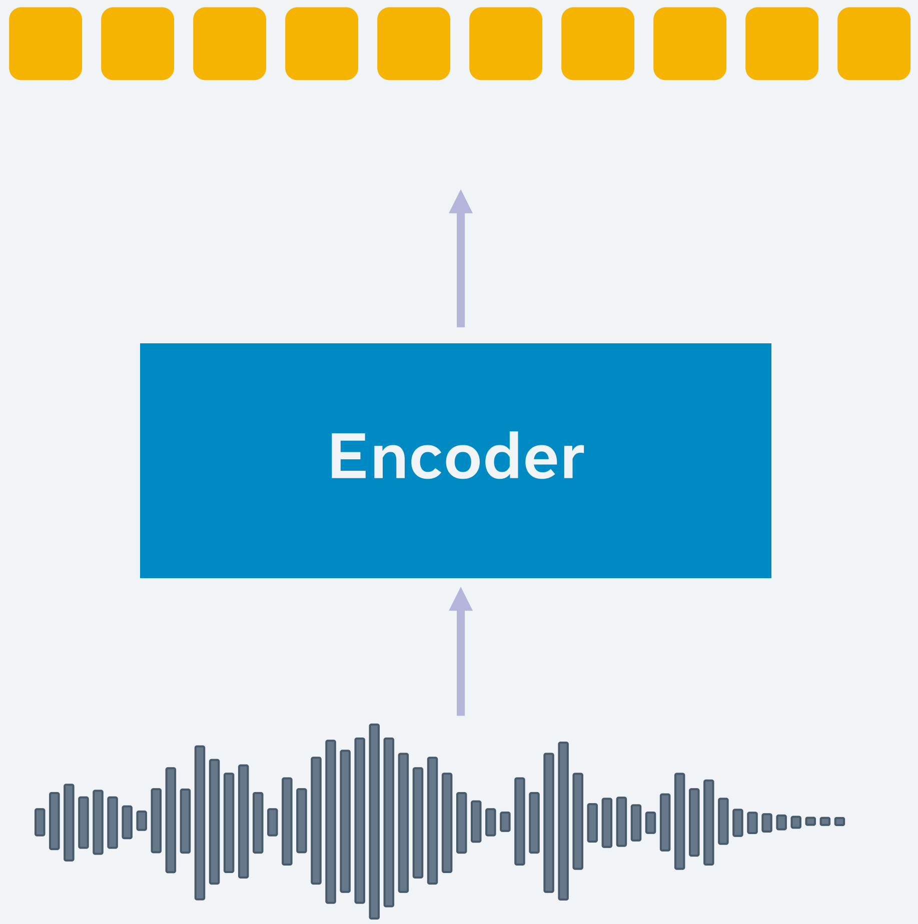
1. CHIMES5. Trmal, Vincent, Watanabe , Barker (2018), Interspeech

Noise robust invariant representations

Data filtering

Data augmentation

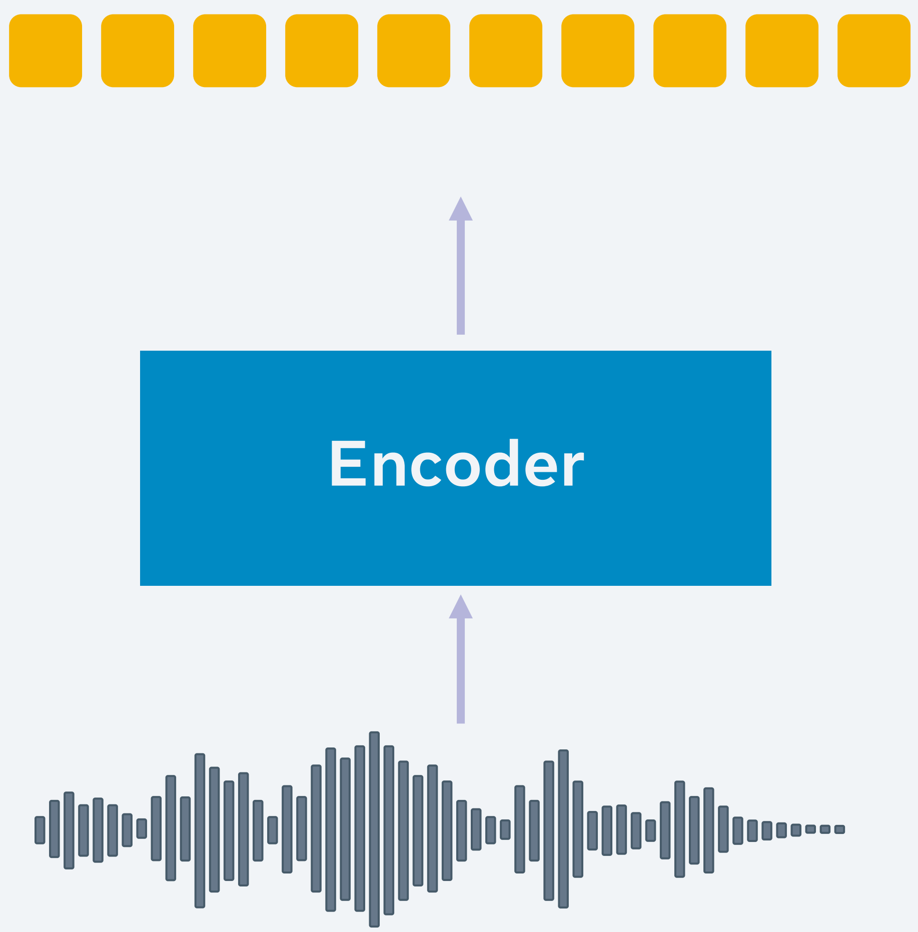
Source separation

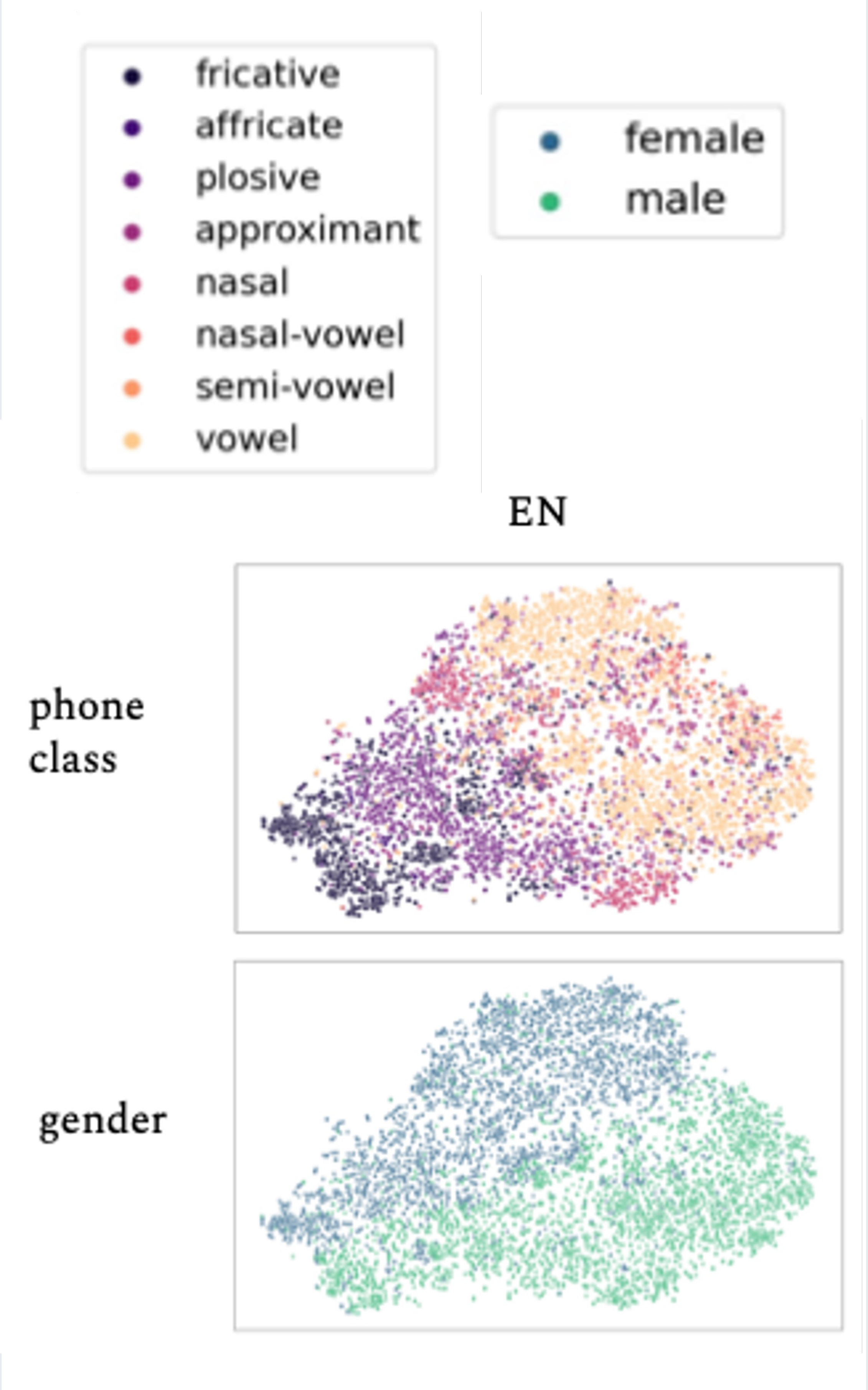


Noise robust invariant representations

Speaker normalisation

Domain adaptation





De Seyssel et al (2022). Interspeech

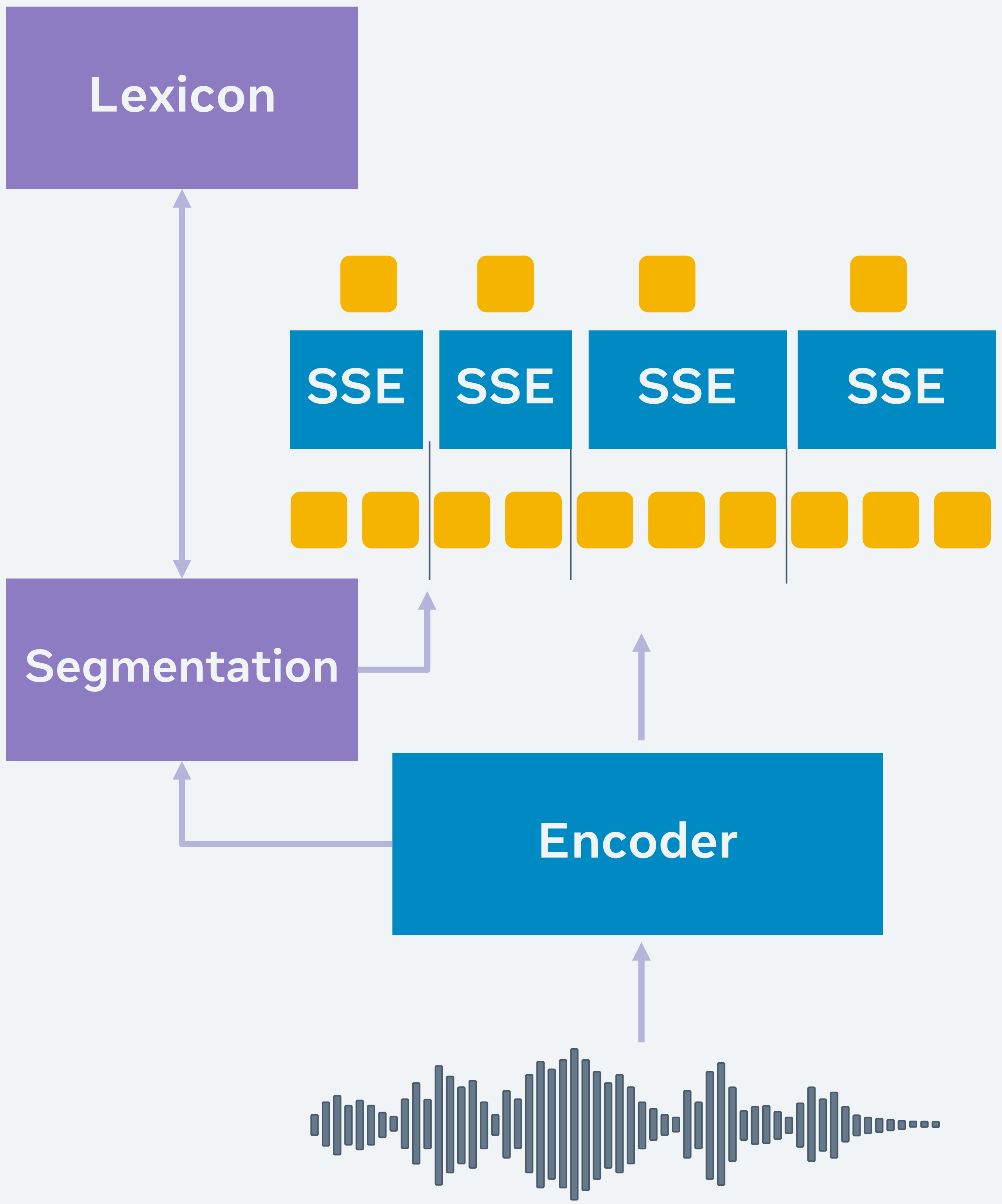
Word discovery

Something is wrong with frame based units!

System	sWUGGY	sBLIMP	sSIMI synth.	sSIMI libri.
CPC-big+km50+BERT-small	65.81	52.91	3.88	5.56
	65.94	53.02	3.02	0.06
CPC-big+km50+LSTM	66.13	53.32	4.42	7.56
	66.22	52.89	7.35	6.66
CPC-small+km50+BERT	70.69	54.26	2.99	6.68
	70.50	54.61	8.96	-1.55
CPC-big+km50+BERT	75.56	56.14	6.25	8.72
	75.51	56.16	5.17	1.75
Forced align BERT	92.19	63.72	7.92	4.54
	91.88	63.16	8.52	2.41
Phone BERT	97.90	66.78	9.86	16.11
	97.67	66.91	12.23	20.16
RoBERTa large	96.58	81.56	32.28	28.96
	96.25	82.11	33.16	27.82

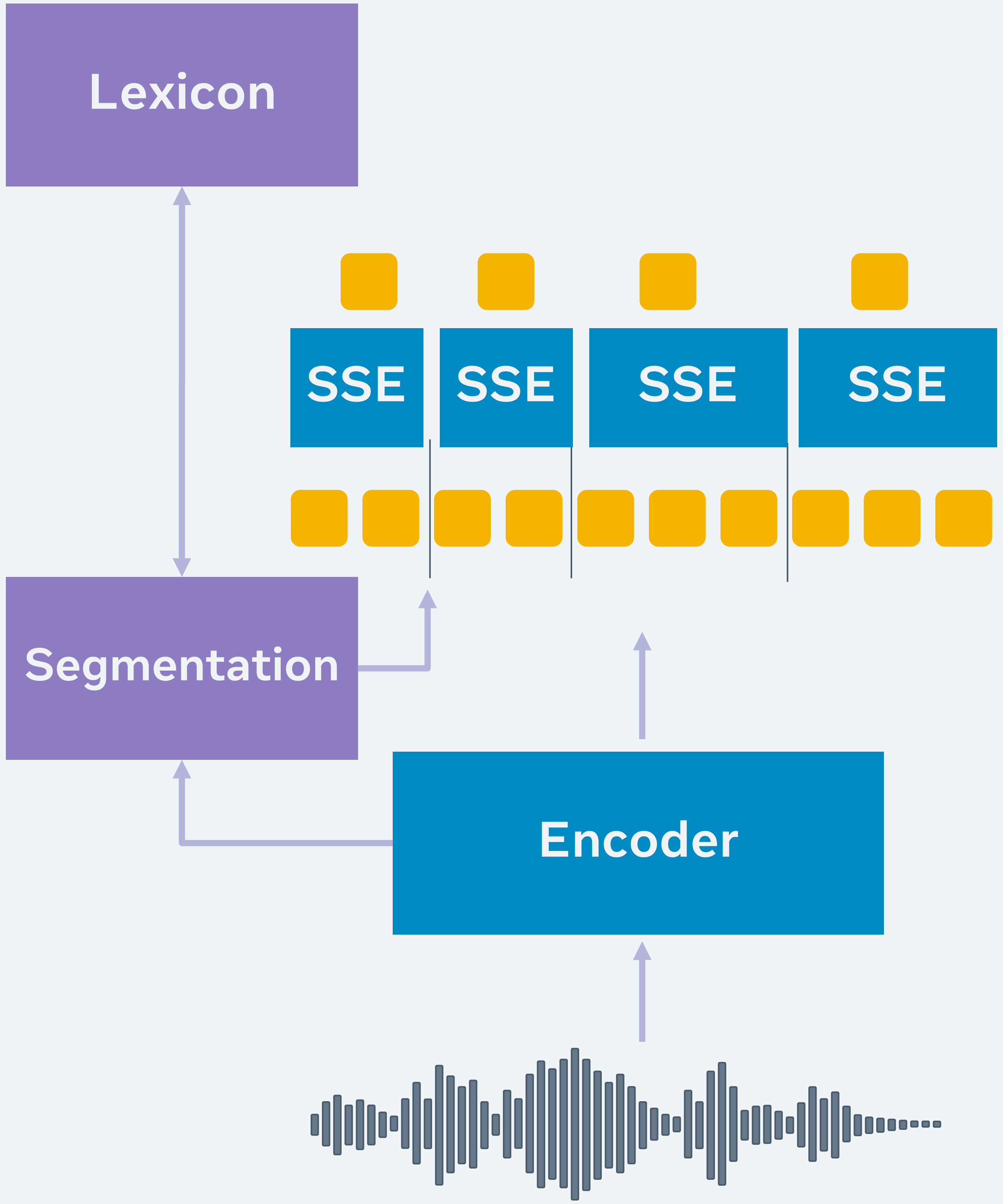
Nguyen et al (2021) The Zero Resource Speech Benchmark 2021: Metrics and baselines for unsupervised spoken language modeling

Word discovery

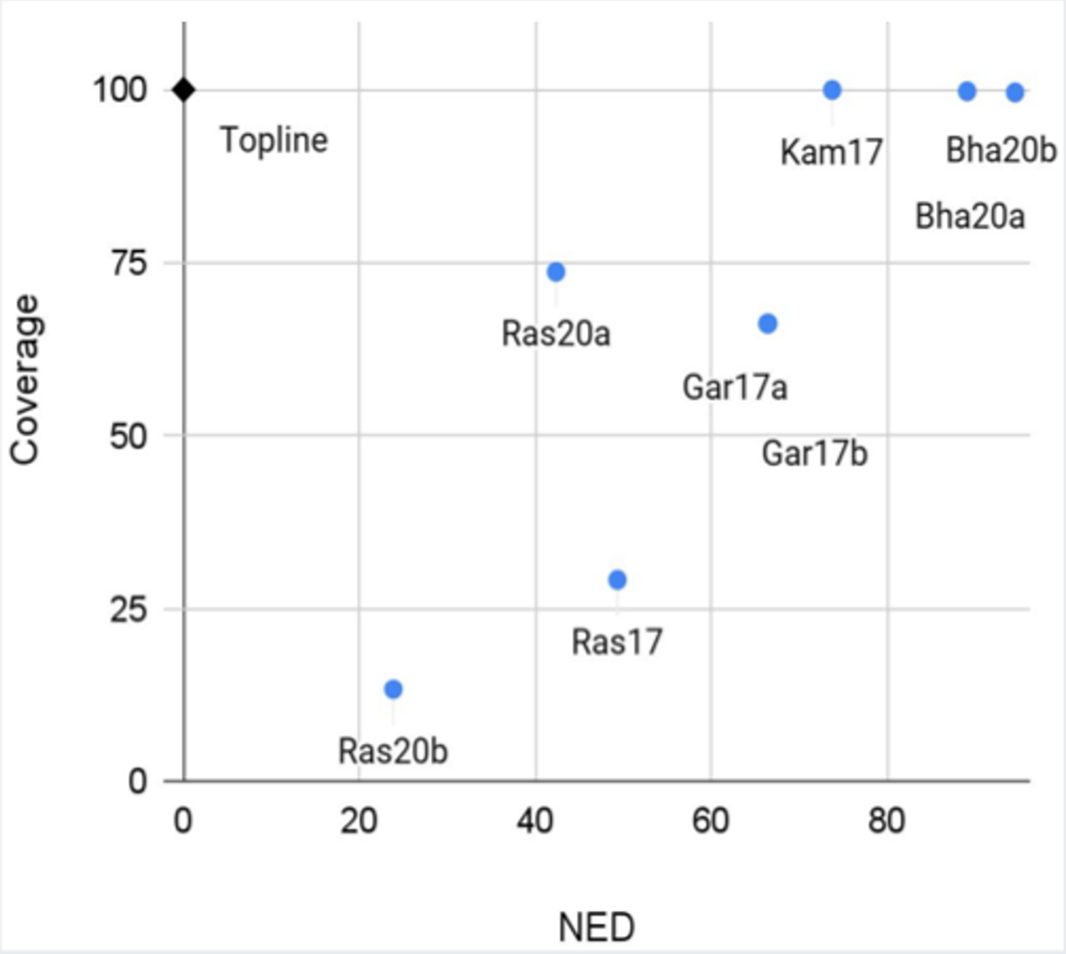


ZRC Task 2
Discover spoken terms and segment with it

Word discovery

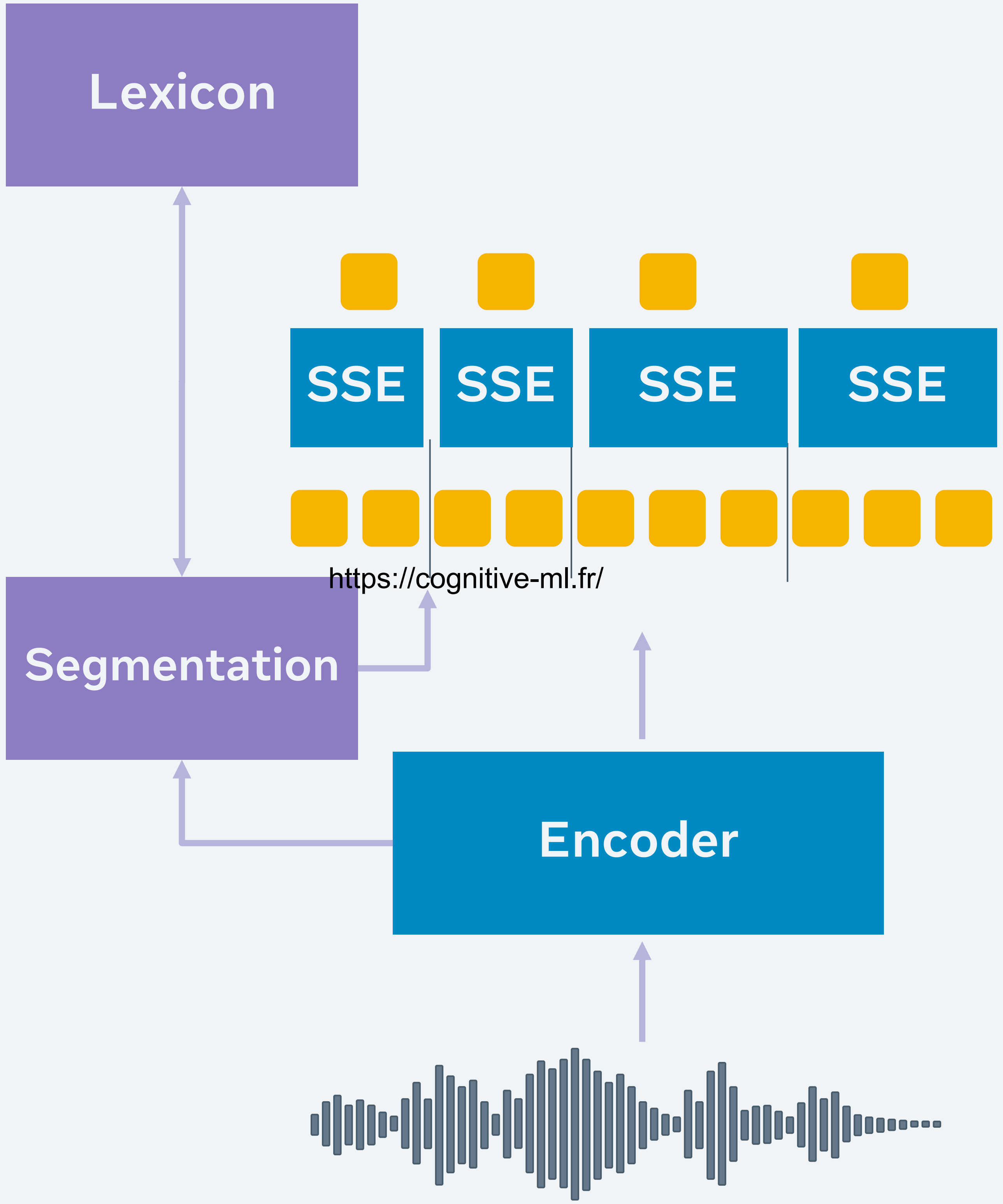


ZRC Task 2
Discover spoken terms and segment with it



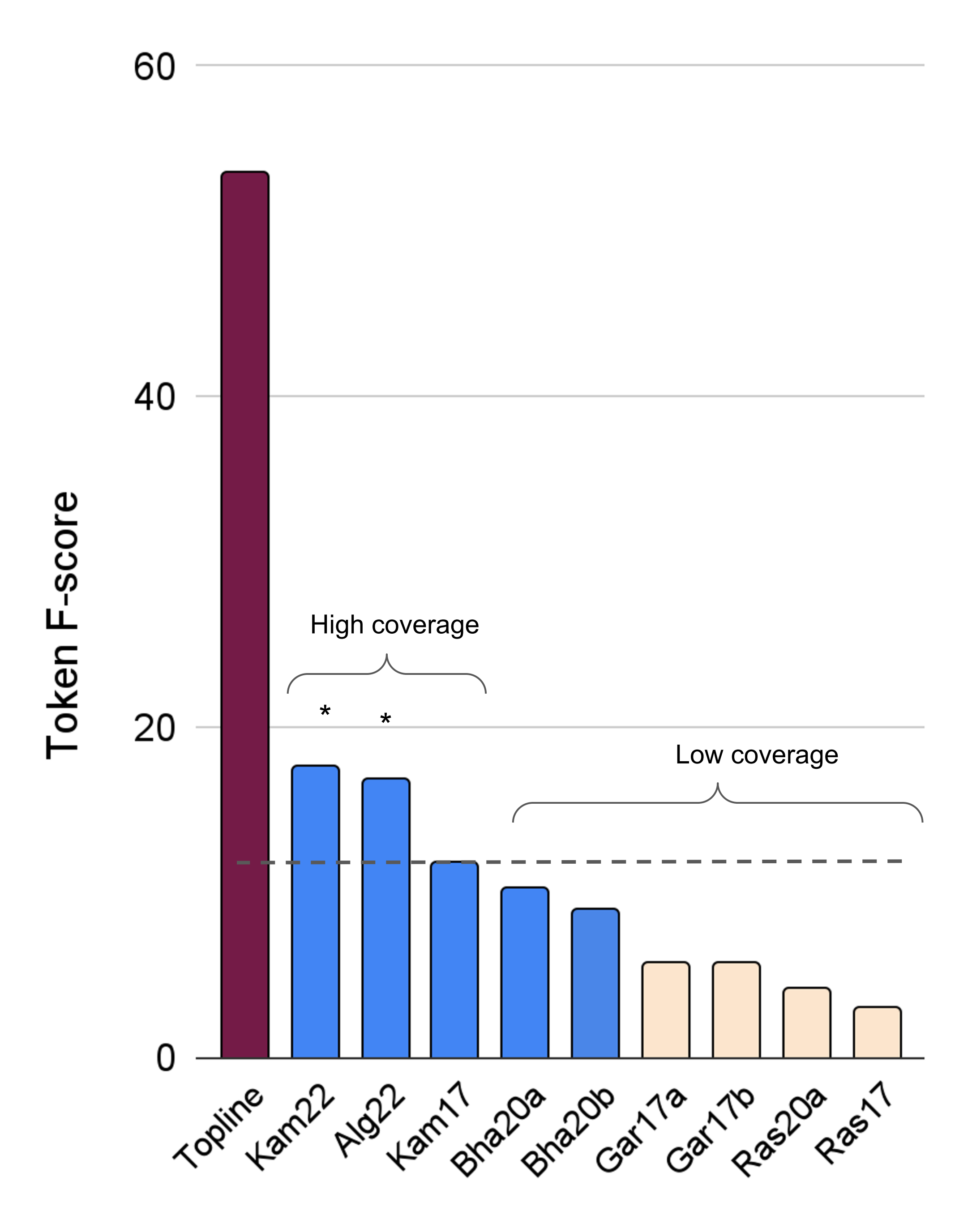
Dunbar, Hamilakis & Dupoux (2022)

Word discovery



ZRC Task 2

Discover spoken terms and segment with it



Dunbar, Hamilakis & Dupoux (2022)

Dataset collection

We need to define a new way to collect expressive speech in a scalable way

LIBRI-LIGHT and VOX POPULI are large (>50k hours) but they are not very expressive.

FISHER, CALLHOME, and other conversational expressive and codatasets already exist but they are too small for current LM training

We propose to collect a new large (100k hours) open source dataset with a smartphone app, allowing different kinds of tasks (debate, story telling, collaborative tasks, open dialogue, etc).

Comments welcome!

Questionnaire for the FISHER 2.0 dataset

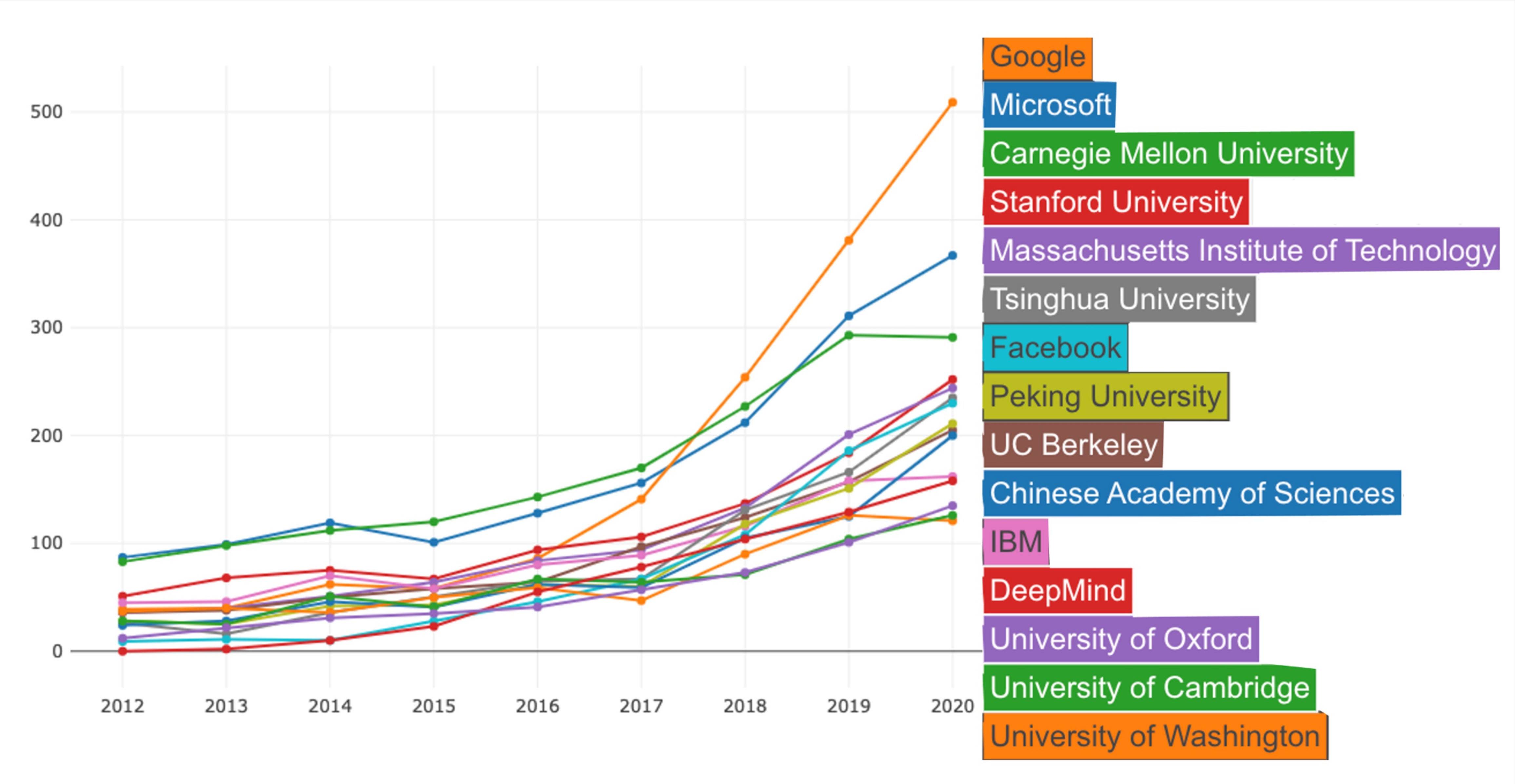
tinyurl.com/3d7d9xmk

Why

Why it matters

TEXT-BASED SERVICES

- Search
- Translate
- Question & Answer
- Recommend
- Describe



Trend of research publications on text-based NLP

Why it matters

SPEECH TO SPEECH SERVICES

- Search
- Translate
- Question & Answer
- Recommend
- Describe

More inclusive

Most languages have no written presence on the web.

More expressive

Intonation, rhythm, sarcasms, laughters, yawning, etc.

More ubiquitous

Online games, local radios, podcasts, metaverse.

Related projects

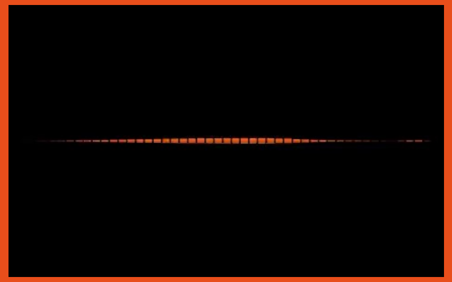
Textless NLP

Self-Supervised Representation Learning

Wav2vec 2.0, HuBERT

Universal Speech Translation

Speech-to-speech translation without text



Datasets and Benchmark

Zero Resource Speech Challenge

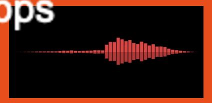
Fisher 2.0

Low bitrate audio compression

Real-time neural codec compressing audio down to 3kbps (100x less than mp3!) beating standard Opus codec at 12kpbs

Opus 6kbps

Encodec 3kbps



Cog/AI: Infant learning simulator

Predicting language development

References

Cognitive machine learning team: https://cognitive-ml.fr/

Questionnaire for the FISHER 2.0 dataset

tinyurl.com/3d7d9xmk

Zero resource speech challenge: Now rolling submissions!

Review paper: Dunbar, Hamilakis & Dupoux (2022). Self-supervised language learning from raw audio: Lessons from the Zero Resource Speech Challenge. JSTSP.

WebSite: www.zerospeech.com

Textless project at Meta

Blog post: https://ai.facebook.com/blog/textless-nlp-generating-expressive-speech-from-raw-audio/ and https://ai.facebook.com/blog/generating-chit-chat-including-laughs-yawns-ums-and-other-nonverbal-cues-from-raw-audio/

Review paper. In progress!

Textless library: https://github.com/facebookresearch/textlesslib

Samples, papers and code: https://speechbot.github.io

Self supervised audio representations

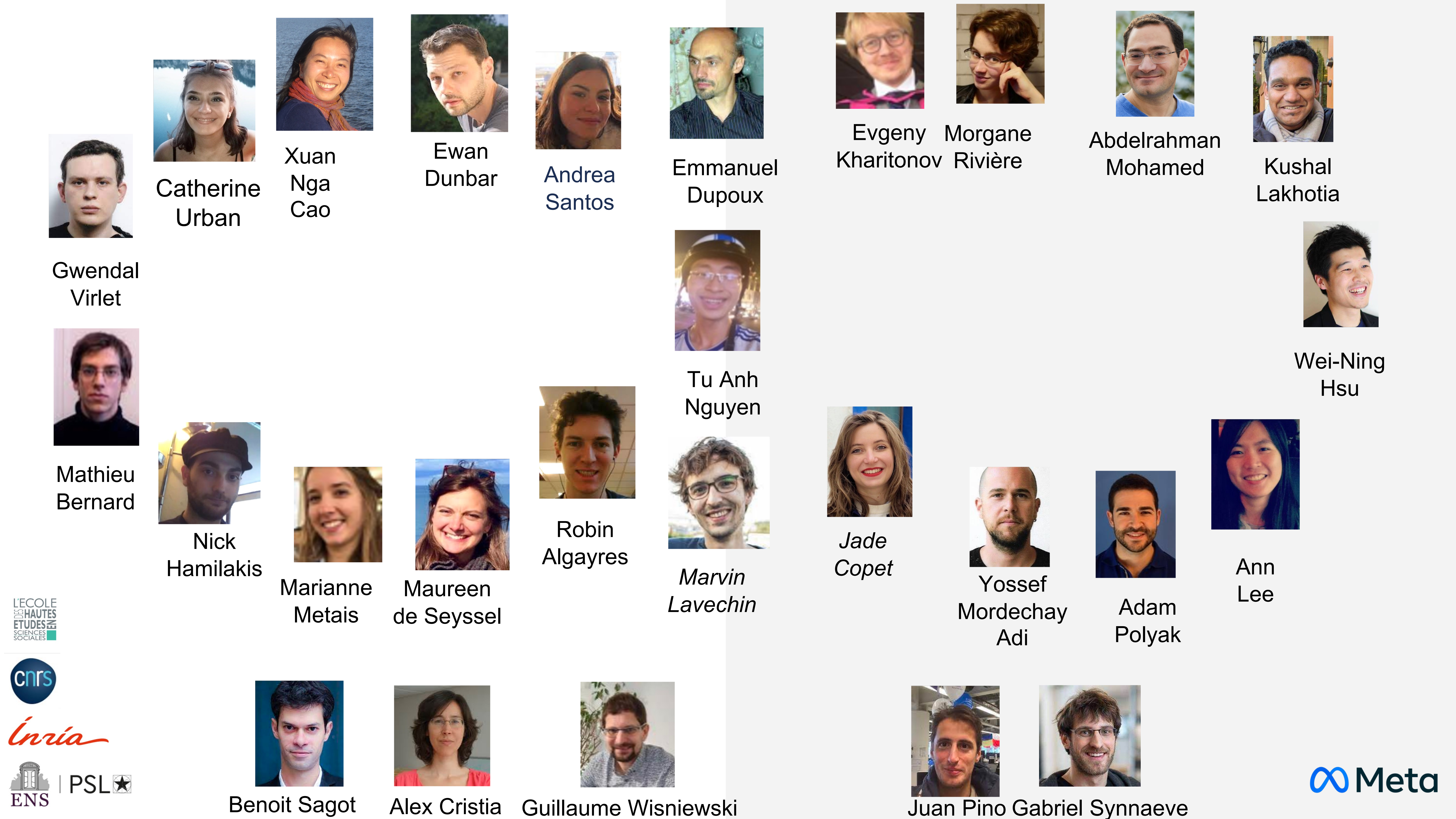
Review paper: https://arxiv.org/abs/2205.10643

Speech to speech translation

Blog Post: https://ai.facebook.com/blog/advancing-direct-speech-to-speech-modeling-with-discrete-units/

New spoken dialogue dataset collection: FISHER 2.0

ongoing


Gwendal Virlet
Catherine Urban
Xuan Nga Cao
Ewan Dunbar
Andrea Santos
Emmanuel Dupoux
Evgeny Kharitonov
Morgane Rivière
Abdelrahman Mohamed
Kushal Lakhotia
Wei-Ning Hsu
Tu Anh Nguyen
Mathieu Bernard
Nick Hamilakis
Marianne Metais
Maureen de Seyssel
Robin Algayres
Marvin Lavechin
Jade Copet
Yossef Mordechay Adi
Adam Polyak
Ann Lee
L'ECOLE DES HAUTES ETUDES EN SCIENCES SOCIALES
cnrs
Inria
ENS | PSL
Benoit Sagot
Alex Cristia
Guillaume Wisniewski
Juan Pino
Gabriel Synnaeve
Meta

Thanks