Uncertainty as a Predictor: Zero-shot MOS Proxies

Aditya Ravuri[UoC] Erica Cooper[NII] Junichi Yamagishi[NII]

Introduction

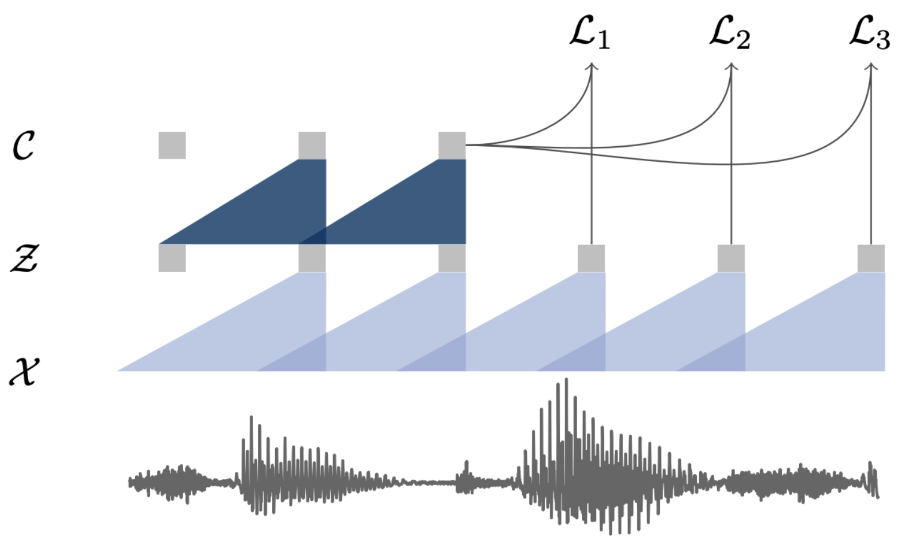
Problem:

- A need for automatic prediction of audio quality (esp. in systhesis and conversion).
- Traditional metrics like Mean Opinion Scores (MOS) challenging to collect at scale & in low-resource settings, and suffer from sampling biases.
- Many prediction solutions use self-supervised (SSL) models as a core component.

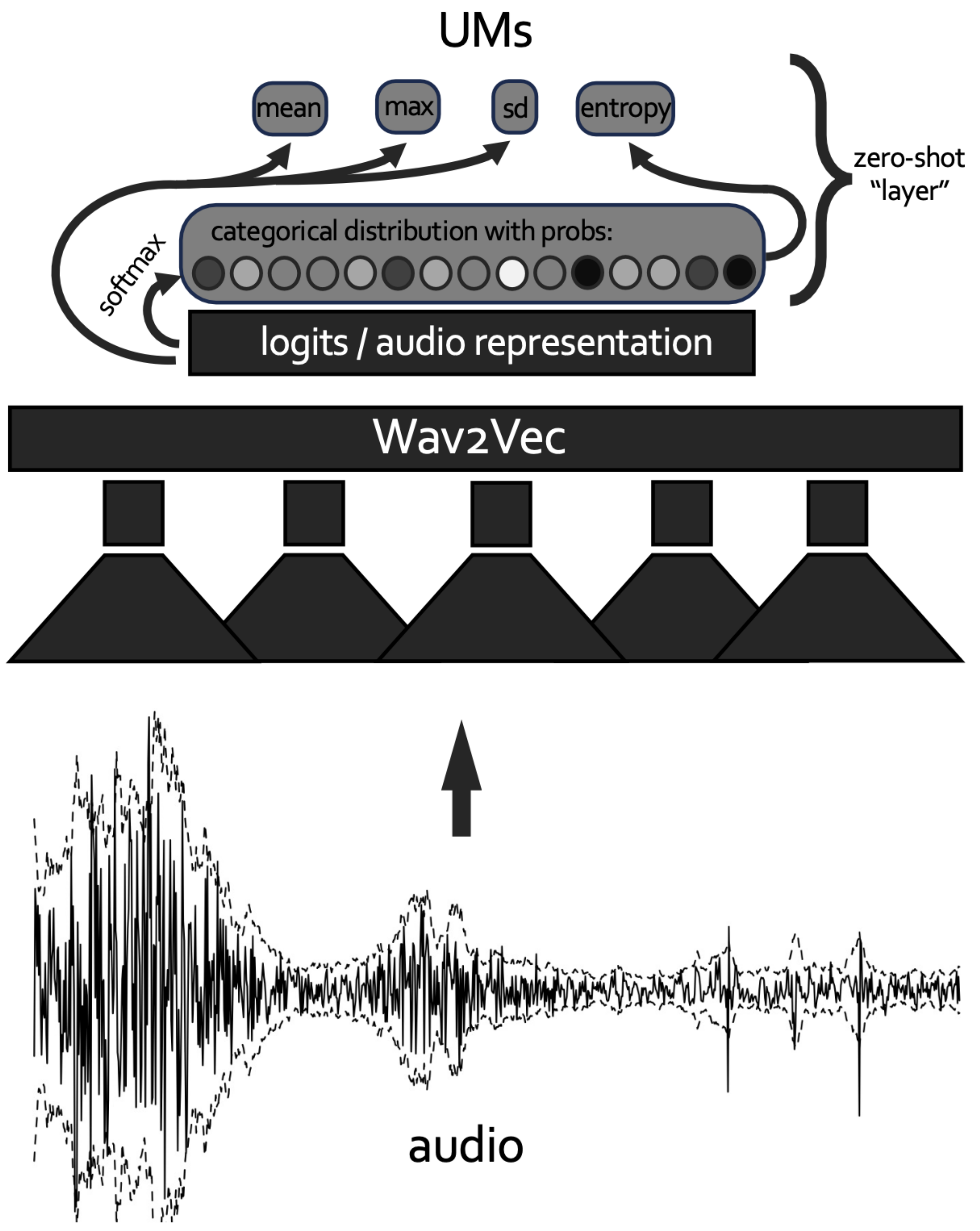
Our Work:

- We show that SSL models such as wav2vec are capable of zero-shot prediction of MOS values.
- We show interesting insights, e.g. that handicapping models can increase their zero-shot performance.

Our hypothesis: SSL models, which were initially used for density estimation, can be used as out-of-data-distribution predictors, through their contrastive probabilities.



Wav2vec's loss is the negative log-probability of predicting a subsequent latent token as opposed to a random token from elsewhere in the audio. Such probabilities were used as data log densities (NCE).


UMs
mean
max
sd
entropy
zero-shot "layer"
categorical distribution with probs:
softmax
logits / audio representation
Wav2Vec
audio

Our approach: we calculate uncertainty measures (e.g. entropy) using an SSL model's contrastive probabilities. If these are high, the model must not be able to distinguish a signal from noise, and so, the signal must be of bad quality.

Main Results

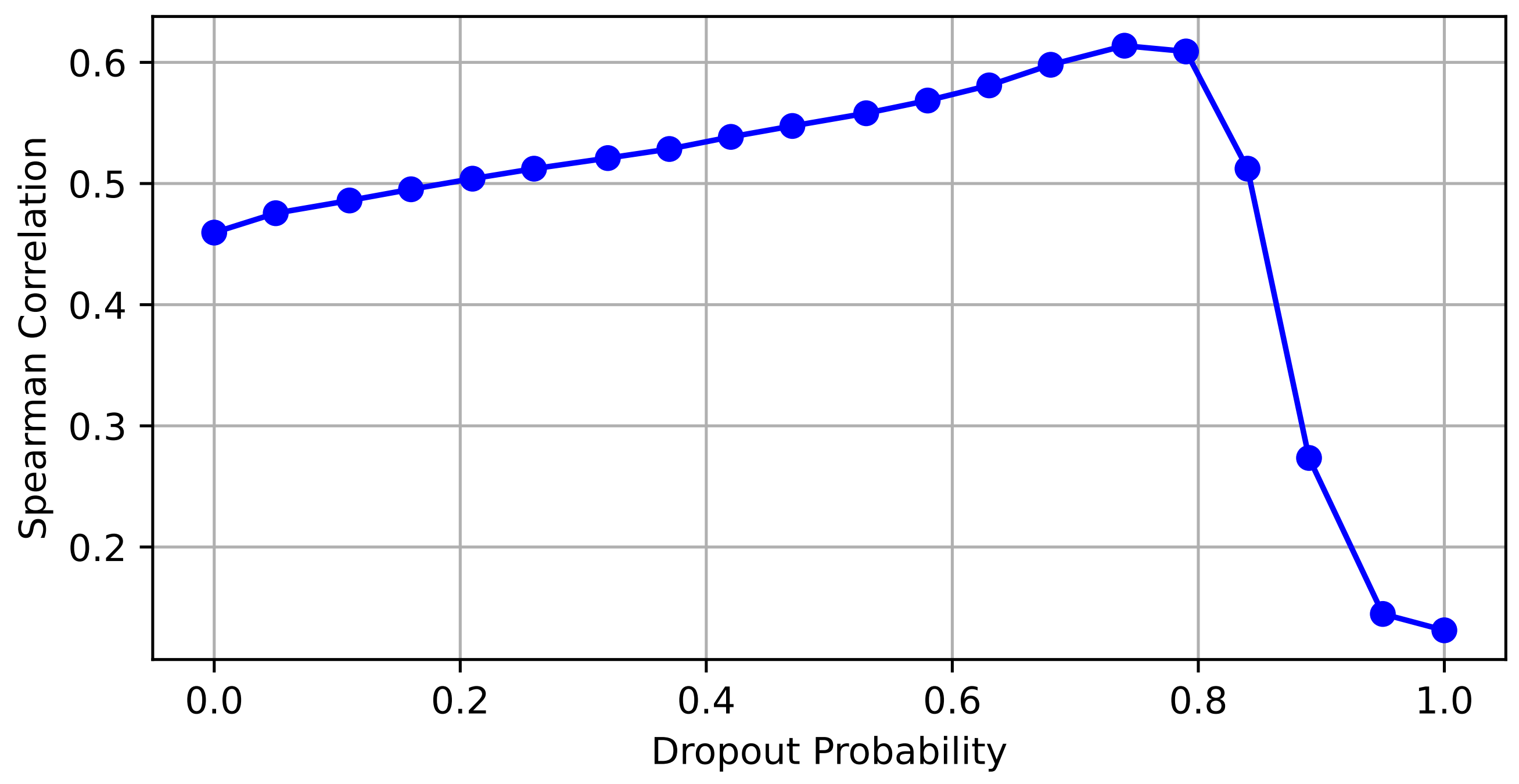
Wav2Vec (and to a lesser extent Wav2Vec2) based UMs correlate highly with MOS values.

SRCC between UMs calculated using various base pre-trained models and MOS values show high correlations when wav2vec is used for zero-shot prediction of MOS values in English settings.

model	type	mean	max	sd	entropy
W2V	Large	**-0.69**	**0.68**	0.64	**-0.70**
	VQ	-0.60	0.58	**0.68**	-0.69
W2V2-ASR	base 960h	-0.46	0.51	0.52	-0.37

Further Insights

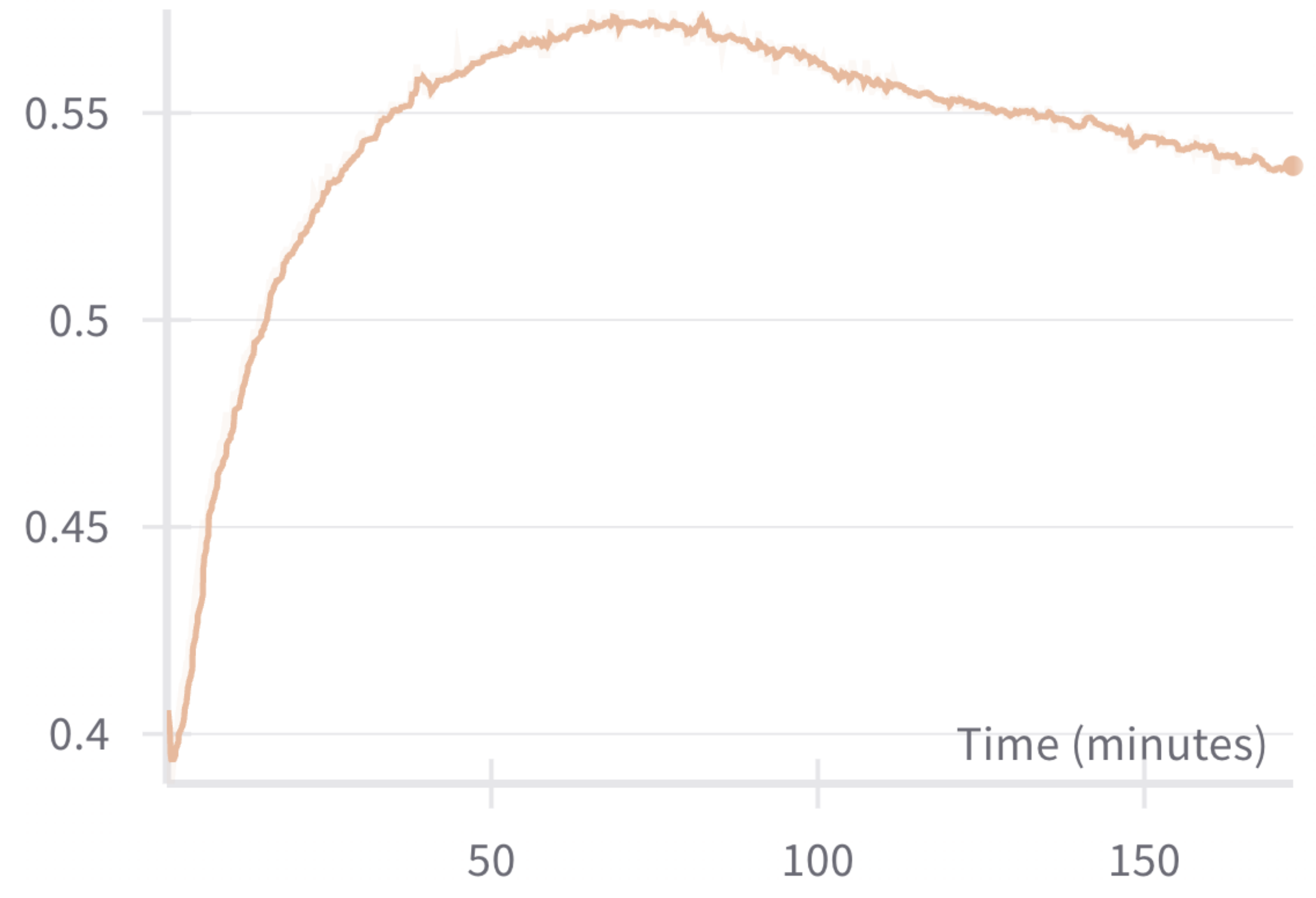
Handicapping wav2vec2 increases performance.



Dropping out features before encoding, and averaging outputs over many forward passes increases performance, by about 15% SRCC when the dropout prob. is around 75%.

UMs are highly correlated with intelligibility measures: transcripts can be generated using a pretrained wav2vec2 base model (which was used to calculate UMs) and an ASR head & language model. UMs share a 70% correlation with word error rates (WERs) computed using these transcripts.

Domain adaptive pretraining increases performance.



Doing a second round of pre-training of the English SSL models on target data of another language (e.g. Mandarin) can significantly increase zero-shot performance on that dataset, although this seems to be dependent on data distribution.

aditya.ravuri@cl.cam.ac.uk

NII UNIVERSITY OF CAMBRIDGE