NOVOIC
Predicting brain disease from speech

Learning de-identified representations of prosody from raw audio

Jack Weston, Raphael Lenain, Udeepa Meepegama, Emil Fristed

What is prosody and why should I care?

NOVOIC

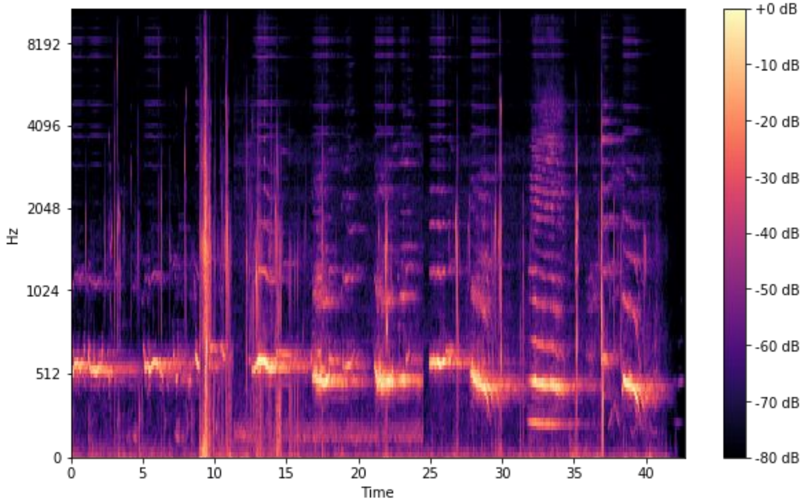


Figure from
https://towardsdatascience.com/getting-to-know-the-mel-spectrogram-31bca3e2d9d0

What is prosody and why should I care?

NOVOIC

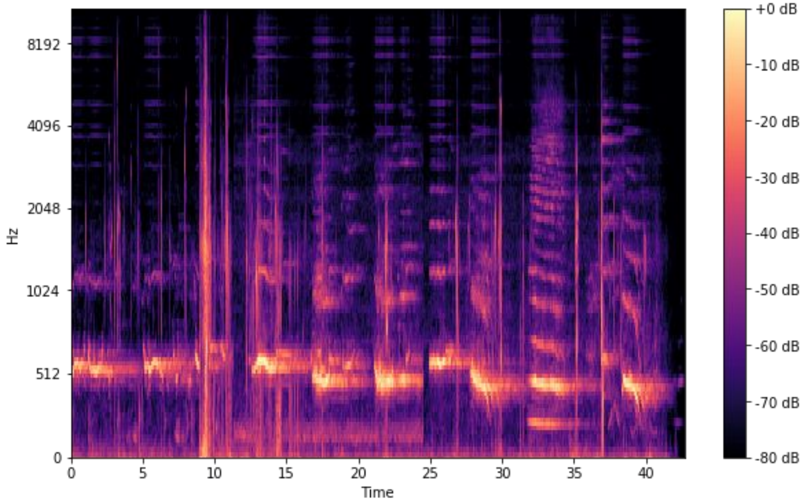


Figure from
https://towardsdatascience.com/getting-to-know-the-mel-spectrogram-31bca3e2d9d0

Generic/mixed

- TRILL on speech
- wav2vec on speech
- Mockingjay on speech
- Averaged spectrogram features
- Crest factor, zero crossing rate etc

Timbral

- TRILL on phonations
- wav2vec on phonations
- Mockingjay on phonations
- Jitter/shimmer
- Acoustic speech features on phonations

Non-timbral

- Speech rate
- Hesitation freq and size
- Pause size

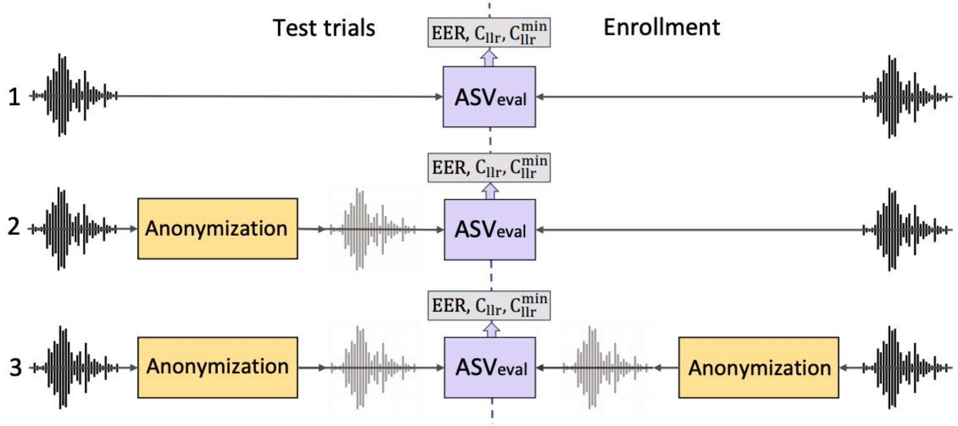
The DAMMP benchmark

NOVOIC

CALC	DAMMP	Dataset	Target	Description	Size	Ref.
-	√	DAIC-WOZ	Depression diagnoses	Interviews by a virtual interviewer	∼300 interviews	(Gratch et al., 2014)
-	√	ADReSS	Alzheimer's disease diagnoses	Picture description tasks	∼200 descriptions	(Luz et al., 2020)
-	√	MUStARD	Sarcasm labels	Acted scenes from TV shows	∼6.4k utterances	(Castro et al., 2019)
√	√	CMU-MOSEI	Sentiment labels	Spoken product reviews	∼20k utterances from ∼2k speakers	(Zadeh et al., 2018)
√	√	POM	Persuasiveness labels	Film reviews	∼300 reviews	(Park et al., 2014)
√	-	TED-LIUM 3	-	TED talks	∼2.4k TED talks	(Hernandez et al., 2018)
√	-	LRS2	-	Single utterances from BBC TV scenes	∼140k utterances	(Afouras et al., 2018)
√	-	AMI	-	Real and acted meetings	∼100 hours of meetings	(Carletta et al., 2005)

Quantifying identifiability of data

NOVOIC



Tomashenko, Natalia, et al. "Introducing the VoicePrivacy initiative." arXiv preprint arXiv:2005.01387 (2020).

$$\mathcal{D}\left(s, r, M(\theta), D\right) := \frac{\mathcal{L}^{online}\left(y_{1:n} | r_{1:n}, s_{1:n}\right)}{\mathcal{L}_{unif}\left(y_{1:n} | r_{1:n}, s_{1:n}\right)} = \frac{t_1}{n} - \frac{1}{n} \sum_{i=1}^{S-1} \log_2 p_{\theta_i}\left(y_{t_i+1:t_{i+1}} | r_{t_i+1:t_{i+1}}, s_{t_i+1:t_{i+1}}\right). \quad (4)$$

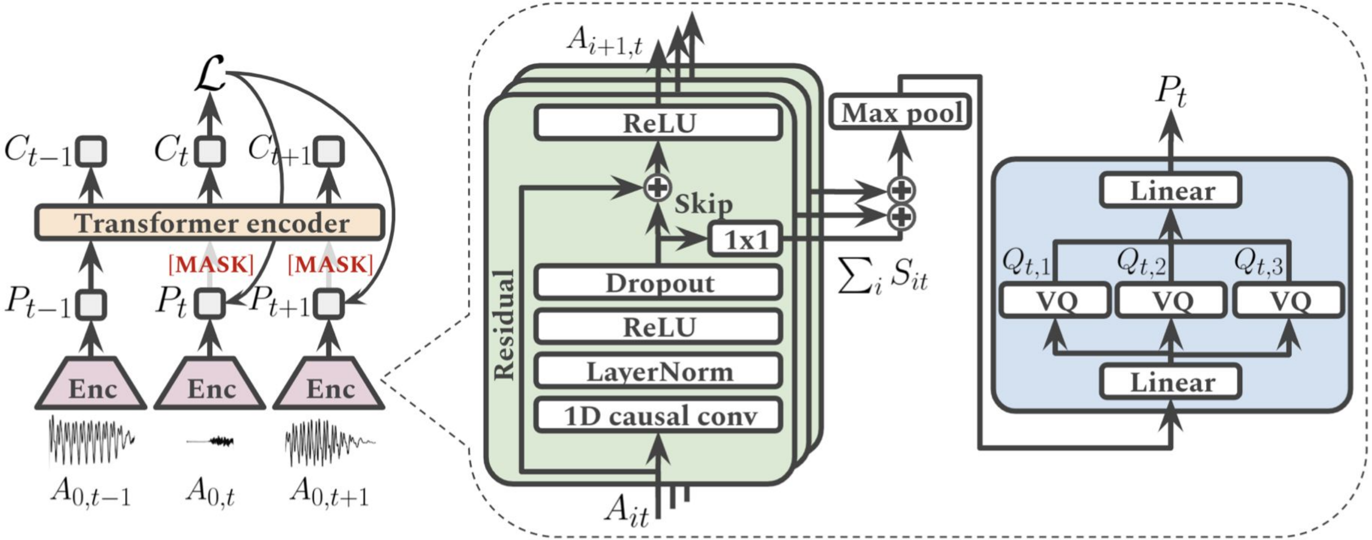
We have that $\mathcal{D} \in [1/n, \infty)$, where $\mathcal{D} = 1/n$ represents the worst possible de-identification and $\mathcal{D} \to \infty$ represents perfect de-identification. This ratio is a function not only of the representations themselves but also of the model $M(\theta)$ and the data set D. As demonstrated in Voita & Titov (2020), the dependence of the codelength on model parameters is relatively light in practice.

Inductive biases

NOVOIC

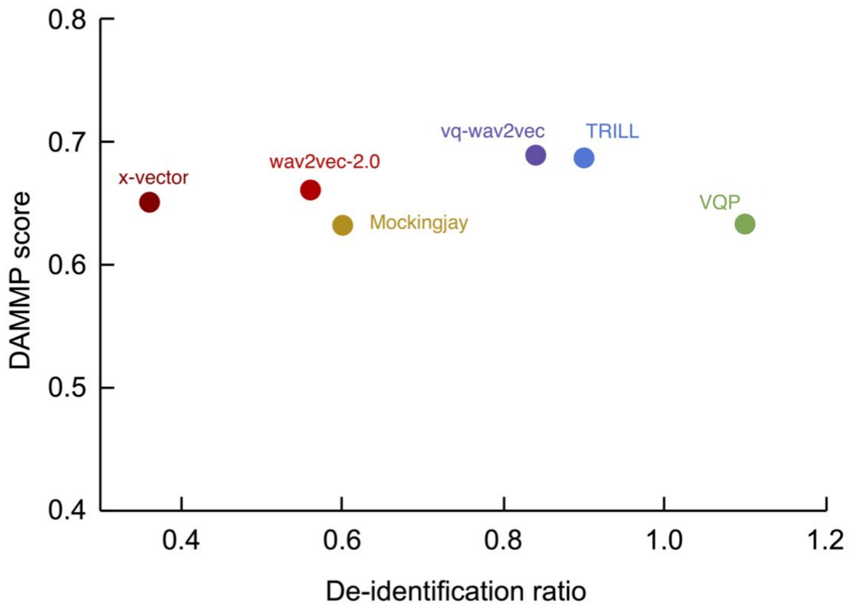
	Inductive bias	What does it do?	Rationale/assumption(s)
Only use audio as input/targets	Prosody itself has predictable temporal patterns.	Learns prosody representations without having to use words/phonemes as input data by relying on predicting temporal patterns requiring strong representations of similar information.	Predicting prosodic states based on prosody alone requires similar prosody representations as predicting prosodic states using words.
Downsample the audio to 500Hz	(Non-timbral) prosody happens <250Hz.	Ensures the network is learning about prosody, not phonetics; makes the input sequence for a word a computationally feasible length.	Nyquist theorem on highest typical female f0 = 2*255Hz =~500Hz
Align the input audio by words; each word learns one prosodic representation	Prosody is strongly temporally associated with/discretized by words.	The prosody encoder creates one independent non-contextualised representation per word.	Semantically meaningful prosody states are naturally discretized on a per-word basis.
Learn vector-quantized representations	There is a finite number of semantically meaningful prosodic states.	Representations must be parsimonious to avoid 'hiding' nuisance covariates in small details => robustness, reliability, generalisation and de-identification.	The most important information for making predictions during self-supervised learning is prosodic.
Contextualization of prosody using e.g. a Transformer encoder	The semantic meaning of prosody is contextual.	Context-aware representations of time-series often make better predictions; contextualization may be the key to disentangling representation from time => audio-linguistic representations.	Contextualisation makes stronger prosody representations for predictions. Contextualization makes prosody representations with weaker cross-temporal interactions, which will help with audio-linguistic representation learning.
Include up to 2s of preceding silence in each audio word	Time between words is part of prosody.	Representations encode information about the absolute/relative speech rate.	Speech rate baseline and temporal variations are an important things to represent. Time preceding is more relevant to the word than time following it.
Use a temporal convolutional network to extract audio features	Prosody is encoded in an audio signal.	Permits a large (1,280 frames) receptive field; learns patterns in periodic signals naturally.	TCNs well-suited to learning patterns in raw audio signals.
Allow ~50*50*50 = 125k quantized prosody states	There is a finite number of semantically meaningful prosodic states.	Expressive enough to represent e.g. 50 semantically meaningful pitches (24 quarter-tones across 2 octaves), 50 semantically meaningful pause lengths and 50 semantically meaningful word rhythms.	125k is enough states to represent most interesting prosody information but not so many that nuisance covariates (e.g. background noise) get represented.

Architecture



Results

NOVOIC



What is VQP representing?

NOVOIC

	TRILL		wav2vec-2.0		vq-wav2vec		Mockingjay		VQP	
	AUC	MDL	AUC	MDL	AUC	MDL	AUC	MDL	AUC	MDL
Pitch										
Pitch	0.558	63.65	0.546	63.88	0.569	63.49	0.558	63.62	0.742	**55.78**
Rhythm										
Intensity	0.596	63.48	0.557	64.19	0.567	64.10	0.558	64.20	0.662	**60.97**
Num. sylls	0.519	65.51	0.508	65.58	0.516	65.48	0.513	65.50	0.616	**63.13**
Tempo										
Artic. rate	0.522	**65.19**	0.506	65.26	0.514	**65.19**	0.510	65.29	0.537	**65.12**
Speech rate	0.532	**64.94**	0.515	65.03	0.519	64.97	0.519	65.01	0.541	**64.88**
Syll duration	0.524	**65.44**	0.509	65.52	0.513	65.48	0.508	65.49	0.497	**65.47**
Word duration	0.544	65.40	0.522	65.58	0.539	65.47	0.536	65.50	0.749	**54.58**
Timbre										
Formant f1	0.735	**58.03**	0.668	62.73	0.696	61.26	0.629	64.07	0.574	65.58
Formant f2	0.743	**57.43**	0.643	63.11	0.666	62.95	0.586	64.87	0.514	65.60
Formant f3	0.779	**54.39**	0.667	62.24	0.688	61.92	0.623	63.90	0.509	65.71

NOVOIC

Predicting brain disease from speech

Jack Weston
jack@novoic.com