Discrete representations in neural models of spoken language

Bertrand Higy
Cognitive Science and AI
Tilburg University
`b.j.r.higy@tilburguniversity.edu`

Lieke Gelderloos
Cognitive Science and AI
Tilburg University
`l.j.gelderloos@tilburguniversity.edu`

Afra Alishahi
Cognitive Science and AI
Tilburg University
`a.alishahi@tilburguniversity.edu`

Grzegorz Chrupała
Cognitive Science and AI
Tilburg University
`grzegorz@chrupala.me`

Abstract

The distributed and continuous representations used by neural networks are at odds with representations employed in linguistics, which are typically symbolic. Vector quantization has been proposed as a way to induce discrete neural representations that are closer in nature to their linguistic counterparts. However, it is not clear which metrics are the best-suited to analyze such discrete representations. We compare the merits of four commonly used metrics in the context of weakly supervised models of spoken language. We compare the results they show when applied to two different models, while systematically studying the effect of the placement and size of the discretization layer. We find that different evaluation regimes can give inconsistent results. While we can attribute them to the properties of the different metrics in most cases, one point of concern remains: the use of minimal pairs of phoneme triples as stimuli disadvantages larger discrete unit inventories, unlike metrics applied to complete utterances. Furthermore, while in general vector quantization induces representations that correlate with units posited in linguistics, the strength of this correlation is only moderate.

1 Introduction

The dominant machine learning paradigm for processing spoken language is based on neural network architectures, such as recurrent nets and transformers, inducing a hierarchy of hidden representations which are distributed and continuous. In contrast, human history has repeatedly seen the discovery and wide adoption of discrete, symbolic representations of speech in the form of writing. These systems commonly represent basic units of language such as morphemes, syllables or phonemes while discarding other information contained in the speech signal such as emotion or speaker identity.

Symbolic representation of speech has proven tremendously useful for storage and transmission of information, and it also plays a crucial role in systems dealing with spoken language such as spoken dialog systems: these typically employ an automatic speech recognition (ASR) module to transcribe the speech signal into written form, which is then used as input to upstream language understanding modules. While some attempts have been made to train such systems end-to-end, pipelines are still very competitive (see Haghani et al., 2018; Higy et al., 2020), which strengthens the point that a symbolic encoding of spoken language contains most of the information relevant for this task.

It may thus be desirable to incorporate similar representations in neural architectures. Accordingly, multiple efforts have been made to design neural networks with discrete hidden representations and to apply them to spoken language data. This is evident in the recent editions of the ZeroSpeech challenge (Dunbar et al., 2019) on unit discovery, which have featured many such approaches. Specifically, Vector Quantization (VQ) has proven to be a simple and effective method to induce discrete neural representations (e.g., van den Oord et al., 2017; Harwath et al., 2020a; Chung et al., 2020; Liu et al., 2021). VQ layers are added to neural architectures in order to map continuous activation vectors onto a finite set of discrete units, often referred to as *codes*, via a dictionary (or *codebook*) associating these codes with their vector embeddings; the number of entries in the codebook is the *codebook size*. Such symbolic codes have been claimed to correspond to phonemes and/or words. What is still lacking though is a detailed analysis of how much the reported equivalence is affected by details of the architectures such as the size and placement of the VQ layers, learning objectives and dataset, as well as by the evaluation metrics used to quantify it. The present study aims

to fill this gap.

We study two approaches to modeling spoken language: learning driven by language-internal structure, and learning driven by grounding in the extra-linguistic world. These two approaches are exemplified by two types of models with VQ layers: the self-supervised model for unit discovery of van Niekerk et al. (2020), and a visually grounded model similar to Harwath et al. (2020a). The datasets used to train each model, Zerospeech 2020 (Dunbar et al., 2020) and Flickr8K (Harwath and Glass, 2015; Rashtchian et al., 2010) are also typical of the task they are used for. Using these two models as our test cases, we systematically investigate the impact of the following factors: (i) the codebook size for the VQ layer, and (ii) the level of placement of the VQ layer. Furthermore, we apply and check the consistency across four different metrics for evaluating the correspondence of the representations with phonemes.

Findings The self-supervised model shows high variability, but with some of the evaluation metrics (especially ABX and RSA, see Section 3.3 for the definition of the metrics) there is a trend for better correspondence to phonemes for smaller codebook sizes (32-128). The visually grounded model, on the other hand, generally shows the closest match with phonemes for the *largest* codebook sizes (512 or 1024) when evaluated on full utterances. In contrast, smaller codebooks score higher for the short utterance segments used by the ABX metric (i.e. minimal pairs of phoneme triples). We also observe inconsistencies in the relative performance of VQ layers placed at different levels in the model for RSA vs. the other metrics. As discussed in Section 5, we attribute those inconsistencies to the properties of the different metrics. Thus, the conclusions drawn using a single metric, even a widely used one like the ZeroSpeech ABX metric, should be treated with caution, and further corroborated.

2 Related work

The domain of speech processing has seen recent efforts to modify existing neural architectures to enable the induction of discrete latent representations. These developments are promising for boosting performance, improving interpretability, and modeling the acquisition and processing of linguistic knowledge in humans.

Applying Vector Quantization (VQ) techniques for this purpose was pioneered by van den Oord et al. (2017), who propose generative models based on the variational auto-encoder (VAE) architecture and use VQ to induce discrete latent representations. They apply this method to images, videos, and speech, and show that the models can learn discrete latent representations without supervision. When applied to raw speech, the VQ-VAE architecture learns high-level discrete representations that are invariant to low-level features of the audio signal such as prosody and speaker identity, and mostly encode the content of the speech. Classification of the discrete representations into phoneme classes (based on majority ground truth label) suggests they capture phonemes to some extent.

Learning discrete instead of (or in addition to) continuous representations can facilitate unit discovery in unsupervised models of speech. In the 2015 Zero Resource Speech Challenge (Versteegh et al., 2015), Badino et al. (2015) present a binarized auto-encoder, certain variants of which outperform its continuous counterpart. In the 2017, 2019, and 2020 editions of this challenge (Dunbar et al., 2017, 2019, 2020), following the work of van den Oord et al. (2017), several models include at least one VQ layer (see e.g. Chorowski et al., 2019; Eloff et al., 2019; Tjandra et al., 2019, 2020). These studies demonstrate the benefit of using VQ layers for phoneme classification and for learning speaker-invariant representations, focusing on the ABX phoneme discrimination metric to evaluate the encoding of phonemic information. However, little analysis on the impact of the size and configuration of the employed VQ layers is performed. For example, van Niekerk et al. (2020) train a VQ-VAE model for reconstructing audio waveforms, and a VQ-CPC (Contrastive Predictive Coding) model for predicting future acoustic units. They evaluate these architectures on the ABX phoneme discrimination task, voice conversion and speaker classification and show that VQ-CPC performs better than VQ-VAE overall, but they do not manipulate the configuration or the dimension of the VQ layers.

Chung et al. (2020), however, report the impact on phoneme and speaker classification of systematic manipulation of VQ-related factors. They train an Autoregressive Predictive Coding (APC) model to predict upcoming frames, and use VQ as a methodology to limit the model's capacity. Using a frame-wise diagnostic classifier (namely linear logistic regression), they show that under restricted configurations (only one VQ layer inserted at the

Table 1: A comparison of studies of VQ-based speech models. RAW (Reconstructing Audio Waveform); VAE (Variational Autoencoder); APC (Autoregressive Predictive Coding); CPC (Contrastive Predictive Coding); DC (Diagnostic Classifier).

Study	Objective	Model	Analysis	Manipulated factors
van den Oord et al. (2017)	RAW	VAE	Phoneme (majority label)	None
Chorowski et al. (2019)	RAW	(V)AE, VQ-VAE	Phoneme (ABX, frame-wise DC), gender, speaker	VAE latent dimensions
Chung et al. (2020)	Predicting next frame	APC	Phoneme (frame-wise DC), speaker	VQ layer number, position, codebook size
van Niekerk et al. (2020)	RAW, predicting future frames	VQ-VAE, VQ-CPC	Phoneme (ABX), speaker, voice conversion	None
Harwath et al. (2020a)	Visual grounding	ResDAVEnet	Phoneme (ABX), bitrate, word (F1 scores)	VQ layer number, position, training regime

top), phoneme prediction using discrete representations improves over using continuous representations learned by an APC model without VQ. In this configuration, larger codebook sizes lead to better performance in phoneme classification but not in speaker identification.

In contrast to the work cited above, which discusses uni-modal speech models, Harwath et al. (2020a) use VQ layers within the setting of learning spoken language via grounding in the visual modality, where the speech signal is associated to images (for an overview of visually grounded speech models, see Chrupała, 2021). They hypothesize that discrete representations learned by such models are more likely to capture higher-level semantic information. Their analyses suggest that the trained multimodal models can learn discrete linguistic units at both word and sub-word levels, with quantization layers inserted at the lower levels of the network showing correspondence to sub-word (i.e. phonemic) units, and those inserted at the higher level corresponding to word-level units. Analyses are based on the Zerospeech ABX metric for phoneme encoding and F1 scores for word detection. Similarly, Liu et al. (2021) propose a framework based on VQ to discover discrete concepts in models of visually grounded language trained on video and text, video and audio or image and audio. Evidence of a correspondence between the learned concepts and visual entities/actions or words are given but no detailed analysis is performed.

When evaluating the encoding of phonemic information in continuous representations from models of spoken language, recent work has shown that different metrics may yield different outcomes. Chrupała et al. (2020) show that representational similarity analysis (RSA) and diagnostic classifier (DC) applied to pooled representations disagree with the results of DC applied on local representa-

tions for RNN-based architectures, while they are all in agreement when applied to transformer-based representations. Algayres et al. (2020) compare the aforementioned ABX metric and the mean-average-precision (MAP) metric (which uses representations to predict whether two stimuli have the same ground-truth label) to each other and to a downstream frequency estimation task. Performance on the three metrics is correlated, but not to a high degree, and marked discrepancies are found for particular models.

Table 1 summarizes some of the representative studies that use VQ layers and their specifications and reported analyses. As can be seen from this summary, existing work on learning VQ-based discrete representations does not easily lead to a coherent picture due to the wide range of the training objectives and modeling architecture they use, the analyses they perform, the evaluation metrics they employ and the VQ-related factors they manipulate. In this paper, we aim to provide this overview by employing different discretized speech modeling approaches and consistently comparing architectural parameters and evaluation metrics.

3 Methods

3.1 Vector quantization

We investigate evaluation metrics for the analysis of discrete representations induced through vector quatization. Since phoneme classification/identification has been the dominant analysis task for discrete representations of speech, we use this as our main task. We do so through the specific case of speech representations learned by two different models: a self-supervised model of speech trained to reconstruct the audio waveform, and a visually-supervised model of spoken language which maps audio representations of spoken ut-

terances and visual representations of their corresponding images to a shared semantic space. Both models employ VQ layers in their architecture to induce discretized representations. A VQ layer takes as input a continuous distributed representation in the form of a vector $h \in \mathbb{R}^d$, and returns the closest of K prototype vectors contained in a trainable codebook $\{e_1, e_2, \ldots, e_K\}$ where $e_i \in \mathbb{R}^d$. For a sequence of continuous vectors $(h_1, h_2, \ldots, h_n)$ the discrete codes are given by the sequence of indices of the prototype vectors returned by the VQ layer. Since the $\arg\max$ operation needed to select the nearest vector is not differentiable, the gradient for backpropagation is approximated by using the straight-through estimator (Bengio et al., 2013), which replaces each non-differentiable operation with the identity function for the backward pass. For further details, consult van den Oord et al. (2017).

3.2 Target models

Self-supervised Our self-supervised model is the VQ-VAE model introduced in van Niekerk et al. (2020).[1] The model consists of an *encoder* built out of a stack of five convolutional layers, a *bottleneck* comprising a linear projection and a VQ layer, and a *decoder* comprising an embedding layer and a stack of upsampling and recurrent layers; the decoder attempts to reconstruct the original waveform. For details of the architecture, see van Niekerk et al. (2020). In the experiments reported here, we vary the size of the codebook, but keep the placement of the VQ layer constant as the encoder contains only one fully connected layer after which the VQ layer can be placed.

Visually-supervised A visually-supervised model of spoken language with discrete representations was introduced by Harwath et al. (2020a): they adapted an existing model (Harwath et al., 2020b) by inserting one or more VQ layers within the speech encoder stack. We similarly adapt the architecture used in Merkx et al. (2019) and Chrupała et al. (2020) by inserting a single VQ layer at one of three levels: either following the first, second or third GRU layer of the speech encoder. In addition to VQ layer placement, we also vary the size of the codebook. Note that unlike Harwath et al. (2020a) we do not use a pre-training stage which by-passes the VQ layers;

---

[1] We use the authors' implementation available at github.com/bshall/ZeroSpeech.

rather, we train the complete network from scratch. This model thus consists of an image encoder, which takes as input image features extracted via a pre-trained ResNet-152 model (He et al., 2016) and maps these features via a learned affine transform into a joint visual-language space. The audio input are MFCC features with total energy and delta and double-delta coefficients with combined size 39. The speech encoder consists of one 1D convolutional layer (with 64 output channels) which subsamples the input by a factor of two, four bidirectional GRU layers (each of size 2048), with a VQ layer inserted between a single pair of GRU layers. This stack is followed by a self-attention-based pooling layer. The objective function is a version of the triplet loss with negative examples from the current batch. The model is trained with the Adam optimizer (Kingma and Ba, 2015) with a cyclical learning rate schedule (Smith, 2017).

3.3 Evaluation methods

Different evaluation methods have been used for analyzing the nature of information captured by VQ-based discrete representations in different studies. We present here a thorough examination of their formal similarities and differences as well as their sensitivity to different conditions. In this section we introduce the methods commonly used to evaluate the learned representations.

Chrupała et al. (2020) study the effect representation scope, i.e. activation vectors retrieved at the level of frames (local) or pooled over whole utterances (global), concluding that it can affect results. Following their recommendations, and for the sake of simplicity, we include one measure for each scope. Since their findings suggest that local RSA lacks sensitivity, we use local DC (which is also the most widely used) as well as global RSA in our experiments. Global RSA has the double advantage of not using any trainable parameters and not requiring any alignments.

Normalized Mutual Information (NMI) An information-theoretically motivated measure of the association between two random variables is mutual information. In the general case of vector-valued neural representations, computing mutual information with the target annotation is intractable. In the special case where the representation is discrete-valued, we can use the standard empirical estimate. Given discrete random variables X

with image $\mathcal{X}$ and Y with image $\mathcal{Y}$ (i.e. frame-wise codes and phoneme labels in our case), the mutual information $I(X; Y)$ is

$$I(X; Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) \log \frac{P(x, y)}{P(x)P(y)} \quad (1)$$

It is often more informative to use mutual information normalized by the arithmetic mean of the entropies of the two random variables:

$$\text{NMI}(X; Y) = 2 \frac{I(X; Y)}{H(X) + H(Y)} \quad (2)$$

where $H(X)$ is the entropy of X. This definition of normalized mutual information (NMI) is equivalent to the V-measure (Rosenberg and Hirschberg, 2007).

Diagnostic Classifier (DC) A diagnostic model, also known as a *probe*, is a classifier or regressor trained to predict some information of interest (such as a linguistic annotation) given a neural representation. To the extent that the model successfully predicts the annotation, we conclude that the neural representation encodes this information. Informally, such a diagnostic classifier can be seen as quantifying the amount of easily-accessible – or in the extreme case, linearly decodable – information about the target annotation (Adi et al., 2017; Alishahi et al., 2017; Hupkes et al., 2018; Conneau et al., 2018, among others).

As argued by Pimentel et al. (2020), without the qualification that information be easily accessible, probing should aim to approximate the mutual information between the neural representation and the target annotation, and thus should use the best-performing probe possible. Furthermore, it is not possible for the neural representation to contain more information about the target annotation than the source utterance itself, due to the information processing inequality, and thus, in the general case, probing with an unrestricted classifier is not a well-founded exercise. In the special case of probing a discrete-valued variable (as is the case in our study) the situation is simpler: the accuracy of a linear classifier is closely related to the empirical estimate of the mutual information between the representation and the target annotation; see the formal argument in Appendix A.1 as well as our empirical results.

Representational Similarity Analysis (RSA) RSA is a second-order technique originating in neuroscience (Kriegeskorte et al., 2008) where similarities between pairs of stimuli are measured in two representation spaces: e.g. neural activation pattern space and a space of symbolic linguistic annotations such as sequences of phonemes or syntax trees. The correlation between these pairwise similarity measurements quantifies how much the two representations are aligned. This approach requires a similarity or distance metric for pairs of stimuli within each representation space, but does not need a way of mapping from one space to the other. It generally does not have any trainable parameters. As a consequence, it is sensitive to the purity of the representation with regard to the variable of interest: unlike DC, the RSA metric will penalize representations for encoding any information unrelated to the target variable. See for example Bouchacourt and Baroni (2018); Chrupała and Alishahi (2019); Abnar et al. (2019); Abdou et al. (2019); Chrupała et al. (2020); Fereidooni et al. (2020); Davis and van Schijndel (2020) for uses in NLP and speech processing.

When the neural representations of the stimuli evaluated are sequences of vectors, we need to make a choice regarding how to measure similarities or distances between them. Here we focus on neural representations which take the form of sequences of symbolic codes, which makes measuring distances simple: a natural choice is the Levenshtein edit distance normalized by the length of the longer string. We can thus apply the same edit-distance metric on both the neural representations and on the reference sequences of phonemes or words (for efficiency we collapsed code repetitions).

ABX discriminability (ABX) The ABX phoneme discriminability metric (Schatz, 2016) as used in the Zerospeech challenge (Dunbar et al., 2019) is based on triples of stimuli (A, B, X) where A and X belong to the same category and B and X belong to different categories. The ABX error is a function of $d(A, X)$ and $d(B, X)$ where $d(\cdot, \cdot)$ is a distance metric for the representation being evaluated:[2]

$$\text{abx}(A, B, X) = \begin{cases} 1 & \text{if } d(A, X) > d(B, X) \\ \frac{1}{2} & \text{if } d(A, X) = d(B, X) \\ 0 & \text{otherwise} \end{cases}$$
$$(3)$$

---

[2]The ABX error is thus similar to a discretized version of the triplet loss.

Metric	Triples	Align	Train	Distance
RSA				✓
ABX	✓	✓		✓
NMI		✓		
DC		✓	✓	

Table 2: Summary of the main features of the evaluation metrics used.

The categories are determined by gold annotation: in the case of Zerospeech they are phoneme labels. The stimuli are presented in context in the form of minimal pairs: (A =/beg/$_1$, B =/bag/, X =/beg/$_2$), where /beg/$_1$ and /beg/$_2$ are two different utterances of this phoneme sequence. In our use case, alignments between the gold phoneme transcriptions and the evaluated representations are required to extract the stimuli for ABX. Here we use the same distance metric as for RSA: Levenshtein edit distance normalized by the length of the longer string.

The ABX error is loosely related to the RSA score. With RSA, pairwise distances are measured between gold representations of stimuli (e.g. their phonemic transcriptions) as well as between system representations of the same set of stimuli. The correlation coefficient between these two sets of distance measurements is the RSA score. With RSA there is no notion of a stimulus triple, but rather the score reflects distances between all pairs of stimuli. Likewise, the representation of stimuli according to the gold standard is typically not in the form of atomic categorical labels but can be any representation with an associated distance (or similarity) metric. Thus RSA can be seen as more general than ABX, while being less controlled.

Summary of metrics Table 2 summarizes the main characteristics of the evaluation metrics described above in the context of analyzing neural representations of speech, along the following facets: the need to arrange input in the form of minimal-paired stimulus triples, the need for alignment between input/codes and phonemic transcriptions, the presence of trainable parameters, and reliance on a distance metric between stimuli. According to these criteria, RSA and NMI are the least restricted in their applicability, requiring only a distance function or alignment, respectively. In addition to a distance metric, ABX needs minimal-paired stimulus triples to be extracted. In addition to alignment DC

has trainable parameters, and in the case of analyzing discrete codes, it behaves like an approximation to the NMI metric.

3.4 Evaluation procedure

We evaluate the induced discrete representations by applying the trained networks on the relevant examples – either full utterances, or speech segments corresponding to sequences of three phonemes (triplets) – and extract the sequences of codes from the VQ layer. We do the same for randomly initialized (untrained) versions of the networks, in order to provide a baseline score, following the methodology of Chrupała et al. (2020). In Section 4, we include in the plots both the baseline scores and the scores with the trained models. For the self-supervised target model we vary only the size of the codebook in the VQ layer,[3] using sizes 2^n for $n \in [5, 10]$. For the visually-supervised target model we use the same sizes and also vary the placement of the VQ layer between one of three levels (following first, second or third GRU layer). For both target models, each variant was trained three times with a different random initialization; the scatter plots in Section 4 show each of these runs, as well as a LOESS fit (Cleveland, 1979) to each combination of codebook size and level.

3.5 Datasets

Training target models Following van Niekerk et al. (2020) we train the self-supervised model on about 15 hours of speech from over 100 speakers provided by Zerospeech 2020 (Dunbar et al., 2020).[4] The visually-supervised model is trained on the Flickr8K Audio Caption Corpus (Harwath and Glass, 2015; Rashtchian et al., 2010)[5], which consists of 8,000 images of daily scenes each paired with five spoken captions. The training portion of this dataset contains 6,000 images and about 34 hours of speech.

Evaluation We encode the development captions of Flickr8K (5,000 captions) using the encoders of the trained and untrained target models, and for each utterance extract the sequence of codes output by the VQ layer. We split this data in half, and use

---

[3]The self-supervised model has the VQ layer as part of its original architecture, in a *bottleneck* composed of only one linear layer making it hard to manipulate the level of the layer without disrupting the model.

[4]https://zerospeech.com/2020/

[5]https://groups.csail.mit.edu/sls/downloads/flickraudio/

one half for training the DC, and the other half for computing the scores for DC, RSA and NMI.

As ABX (and one experiment with RSA) is not computed on full utterances but on phoneme trigrams, we prepare this data by sampling 1,000 captions from the Flickr8K development set, and cutting the audio into non-overlapping segments corresponding to a sequence of three phonemes. We then use the ZeroSpeech code to generate minimal-pair stimulus triples.

In order to obtain reference phonemic transcriptions we use forced alignment with the Gentle toolkit,[6] based on Kaldi (Povey et al., 2011). This fails for a small number of utterances, which we remove from the data.

3.6 Repository

The code for replicating our experiments is available at https://github.com/bhigy/discrete-repr under Apache License 2.0.

4 Experimental results

Here we report experiments examining how VQ layers encode phonemes in each target model according to different evaluation metrics. The impact of VQ layers on performance of the visually-supervised model in image retrieval is reported in Section A.2 of the Supplementary Material.

4.1 Visually-supervised representations

We extract codes from visually-supervised VQ models trained on the Flickr8K data while varying VQ layer placement and codebook size. We evaluate how much these codes correspond to phonemes according to four metrics: DC, NMI, RSA and ABX. These results are shown in Figure 1.

DC vs. NMI As expected on theoretical grounds, diagnostic classifiers and mutual information give very similar results. Overall, VQ layers at level 1 and 2 perform significantly better than VQ layers at level 3 or equivalent untrained models. Larger codebook sizes also tend to perform better than smaller ones.

DC and NMI vs. RSA RSA differs from the other two metrics on three main aspects: (i) VQ layers at level 3 perform comparably to level 2, (ii) medium codebook sizes give better performances for VQ layers at level 1, and (iii) untrained models show very poor performance. The three points

[6]http://lowerquality.com/gentle/

Model	RSA input	r
VS	triplet	0.93
VS	complete	0.16
SS	triplet	0.92
SS	complete	0.93

Table 3: Correlation between the ABX and RSA scores, with RSA computed on complete utterances and phoneme triplets, for the visually supervised (VS) and self-supervised (SS) models.

can be explained by the sensitivity of RSA to the purity of the representations. Focusing on the last point first, while representations extracted from untrained models can still contain meaningful information, it will be less explicit and mixed with information that is not relevant for the task of interest. A similar explanation might also hold for the first two points; in particular, we observed that the VQ layers at level 1 retain much more information on the speaker than the other two levels and that the encoding of speaker information increases with the size of the codebook (see Section A.4 of the Supplementary Material for related experiments).

ABX vs. rest ABX is the most divergent metric. While VQ layers at level 1 and 2 still perform better, the gap with equivalent untrained models is smaller. The effect of codebook size is also much less pronounced and layer-specific. The difference in patterns of results between the metrics might be due to the different testing stimuli used by ABX versus the other three metrics: whereas DC, NMI and RSA are tested on full utterance audio files, ABX is tested on small phoneme trigram files.

Role of stimulus size To disentangle the impact of the metric and the size of the stimulus it is applied on, we run two additional experiments. First, we re-calculate RSA using the same phoneme trigram files used for ABX. The correlation coefficient between ABX and this version of RSA is much stronger (see Table 3), suggesting that the type of stimulus used to test the model does play a role.

The other experiment goes in the opposite direction and brings the ABX evaluation closer to the other three metrics. Training the target model on full sentences but applying it to short segments could play a role. Thus, we run an additional set of experiments where we apply the models to full utterances and generate the representation used in
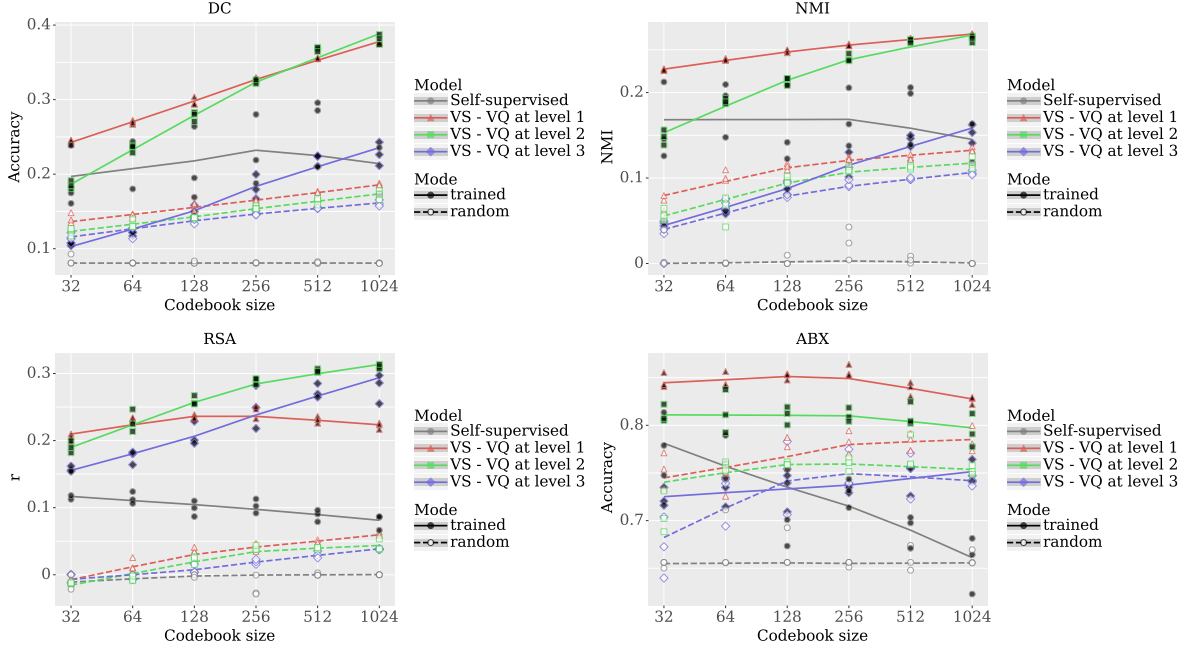
Figure 1: Correspondence of codes to phonemes according to four different metrics as a function of codebook size, type of model, and level of VQ layer placement (in the visually-supervised case).
Top left: Accuracy of the DC.
Top right: NMI between codes and phoneme labels.
Bottom left: RSA score measured against phonemic transcriptions.
Bottom right: Accuracy on the ABX task.
RSA is computed on full utterances; ABX on phoneme-triple segments. DC and NMI are computed frame-wise on full utterances forced-aligned to the phonemic transcriptions. Higher scores are better for all metrics. Trained models are represented by black icons and solid lines. Dashed lines and white filled icons represent randomly initialized models.

ABX by extracting the portion of the code sequence corresponding to each phoneme trigram from the full sequence of activations. The correlation with RSA is still low (0.14) indicating that the problem is intrisic to the evaluation relying on phoneme triplets and not train/test mismatch. This impact of stimulus size on results is likely related to the fact that, with very short stimuli, most normalized edit distances will be maximum, or near maximum, and this will especially be the case for large codebook sizes, giving very skewed and long-tailed edit distance distributions (see details in Section A.3 in the Supplementary Material).

4.2 Self-supervised representations

We extract codes from a number of self-supervised VQ-VAE models trained on the Zerospeech 2019 challenge dataset with varying codebook sizes (note that the VAE model has only one possible placement for the VQ layer). We compare how well these codes correspond to phonemes according to the same four metrics, as displayed in Figure 1. In

contrast to what we see for visually-supervised representations, here RSA and ABX scores are largely consistent and suggest that a larger codebook leads to weaker encoding of phonemic information. The effect is more pronounced for ABX though with the largest codebook performing similarly to the baseline. DC and NMI are relatively insensitive to the size of the codebook.

The self-supervised model does not show the discrepancy observed with the visually-supervised model when RSA and ABX scores are run on testing input of different size. This is confirmed by the correlation coefficients between the ABX score and RSA computed on complete utterances and phoneme triplets shown in Table 3. The VQ layer in the self-supervised model only has access to a limited context, which provides enough information to subsequently reconstruct the input audio frame. The visually-supervised architecture on the other hand builds a representation for the whole utterance, supported by recurrent layers. This core difference may explain the pattern of results we

report here.

5 Discussion

To summarize, the different metrics we compared give divergent views of the same representations. This should not necessarily be interpreted as one metric being right while the others are flaweded. It is likely that the different metrics account for somewhat different properties of the representation. The differences between RSA and DC/NMI are probably related to the purity of the representation. RSA is based on the correlation of distances between pairs of stimuli and is thus sensitive to the presence of additional information in the representation. If the model's representation contains information that is not related to the representation that is the target of the analysis (in our case phonemes identification), this will be reflected in lower RSA scores.

An obvious example of such information that a model of speech is likely to encode is speaker identity. As pointed out in Section 4.1, the pattern of results obtained with a classifier trained to predict speaker identity supports this view. The encoding of speaker information could lead to comparatively lower scores with RSA, especially for level 1 and large codebooks where speaker identity is better encoded.

In general, our results suggest that different metrics might be preferred depending on the question that one is trying to answer. RSA scores are a better indicator of the exact match between two representations while DC/NMI better evaluate the extent to which a given information can be extracted from a model's representation, irrespective of other sources of information that might be encoded at the same time. This could be confirmed through white-box experiments, where the different metrics would be applied to hand-crafted representations with different properties (e.g. in term of purity). We leave this to future work.

The only remaining point of concern that arises from our results is the interaction between codebook size and input size. Larger codebook sizes tend to be disadvantaged when short input segments are used as input, such as minimal pairs of phoneme triplets. This is particularly relevant for ABX where the use of minimal pairs of phoneme triplets is a common practice. Analysis of discrete representations using short segments should preferably be carried with diagnostic classifiers or NMI,

especially if codebooks or different sizes are compared.

It is also important to highlight that the importance of this effect is dependent on the architecture and the training objective of the model, as our experiments with a self-supervised model show.

6 Conclusion

We compared four different metrics for discrete representations induced by VQ layers in weakly-supervised models of spoken language, and while the results are broadly consistent, some differences did emerge. RSA tends to show a bigger gap between trained and untrained models as it is more sensitive to the purity of representations with respect to the information of interest. More surprising is the divergent results we observe when evaluation is performed on minimal pairs of phoneme trigrams: this is likely due to the skew of distance distributions with large codebooks sizes.

This is an important finding as some previous work on discrete representations focused exclusively on the ABX metric. In contrast, we recommend corroborating results with multiple analytical approaches, as currently their behavior in different settings is incompletely understood.

Overall, our findings do support the idea that vector quantization is an effective way to induce discrete representations, and that these correlate with symbolic representations assumed in linguistics. However, it is worth noting that the absolute values of our metrics measuring correspondence to phonemes are moderate at best.[7] It is thus important to keep in mind that these symbolic units are not exact analogs of the concepts familiar from linguistic theory and psycholinguistic studies.

Acknowledgements

References

Mostafa Abdou, Artur Kulmizev, Felix Hill, Daniel M. Low, and Anders Søgaard. 2019. Higher-order comparisons of sentence encoder representations. In

---

[7]The highest RSA correlations are around 0.3 and NMI values slightly above 0.25.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5838–5845, Hong Kong, China. Association for Computational Linguistics.

Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. 2019. Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 191–203, Florence, Italy. Association for Computational Linguistics.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Robin Algayres, Mohamed Salah Zaiem, Benoît Sagot, and Emmanuel Dupoux. 2020. Evaluating the reliability of acoustic speech embeddings. In *INTERSPEECH 2020 - Annual Conference of the International Speech Communication Association*, Shanghai / Vitrtual, China.

Afra Alishahi, Marie Barking, and Grzegorz Chrupała. 2017. Encoding of phonology in a recurrent neural model of grounded speech. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 368–378, Vancouver, Canada. Association for Computational Linguistics.

Leonardo Badino, Alessio Mereta, and Lorenzo Rosasco. 2015. Discovering discrete subword units with binarized autoencoders and hidden-markov-model encoders. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

Diane Bouchacourt and Marco Baroni. 2018. How agents see things: On visual representations in an emergent language game. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 981–985, Brussels, Belgium. Association for Computational Linguistics.

Jan Chorowski, Ron J. Weiss, Samy Bengio, and Aäron van den Oord. 2019. Unsupervised Speech Representation Learning Using WaveNet Autoencoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12):2041–2053. Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Grzegorz Chrupała and Afra Alishahi. 2019. Correlating neural and symbolic representations of language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2952–2962, Florence, Italy. Association for Computational Linguistics.

Grzegorz Chrupała, Bertrand Higy, and Afra Alishahi. 2020. Analyzing analytical methods: The case of phonology in neural models of spoken language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4146–4156, Online. Association for Computational Linguistics.

Grzegorz Chrupała. 2021. Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques. *CoRR*, abs/2104.13225.

Yu-An Chung, Hao Tang, and James Glass. 2020. Vector-quantized autoregressive predictive coding. *arXiv preprint arXiv:2005.08392*.

William S Cleveland. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Forrest Davis and Marten van Schijndel. 2020. Discourse structure interacts with reference but not syntax in neural language models. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 396–407, Online. Association for Computational Linguistics.

E. Dunbar, X. N. Cao, J. Benjumea, J. Karadayi, M. Bernard, L. Besacier, X. Anguera, and E. Dupoux. 2017. The zero resource speech challenge 2017. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 323–330.

Ewan Dunbar, Robin Algayres, Julien Karadayi, Mathieu Bernard, Juan Benjumea, Xuan-Nga Cao, Lucie Miskic, Charlotte Dugrain, Lucas Ondel, Alan W Black, et al. 2019. The zero resource speech challenge 2019: Tts without t. *arXiv preprint arXiv:1904.11469*.

Ewan Dunbar, Julien Karadayi, Mathieu Bernard, Xuan-Nga Cao, Robin Algayres, Lucas Ondel, Laurent Besacier, Sakriani Sakti, and Emmanuel Dupoux. 2020. The Zero Resource Speech Challenge 2020: Discovering Discrete Subword and Word Units. In *Proc. Interspeech 2020*, pages 4831–4835.

Ryan Eloff, André Nortje, Benjamin van Niekerk, Avashna Govender, Leanne Nortje, Arnu Pretorius, Elan van Biljon, Ewald van der Westhuizen, Lisa van Staden, and Herman Kamper. 2019. Unsupervised acoustic unit discovery for speech synthesis using discrete latent-variable neural networks. *arXiv:1904.07556 [cs, eess]*. ArXiv: 1904.07556.

Sam Fereidooni, Viola Mocz, Dragomir Radev, and Marvin Chun. 2020. Understanding and improving word embeddings through a neuroscientific lens. *bioRxiv*.

Parisa Haghani, Arun Narayanan, Michiel Bacchiani, Galen Chuang, Neeraj Gaur, Pedro Moreno, Rohit Prabhavalkar, Zhongdi Qu, and Austin Waters. 2018. From Audio to Semantics: Approaches to End-to-End Spoken Language Understanding. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 720–726. ISSN: null.

David Harwath and James Glass. 2015. Deep multimodal semantic embeddings for speech and images. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 237–244. IEEE.

David Harwath, Wei-Ning Hsu, and James R. Glass. 2020a. Learning hierarchical discrete linguistic units from visually-grounded speech. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. 2020b. Jointly discovering visual objects and spoken words from raw sensory input. *International Journal of Computer Vision*, 128(3):620–641.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Bertrand Higy, Desmond Elliott, and Grzegorz Chrupała. 2020. Textual Supervision for Visually Grounded Spoken Language Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2698–2709, Online. Association for Computational Linguistics.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4.

Alexander H. Liu, SouYoung Jin, Cheng-I. Jeff Lai, Andrew Rouditchenko, Aude Oliva, and James Glass. 2021. Cross-Modal Discrete Representation Learning. *arXiv:2106.05438 [cs]*. ArXiv: 2106.05438.

Danny Merkx, Stefan L. Frank, and Mirjam Ernestus. 2019. Language Learning Using Speech to Image Retrieval. In *Proc. Interspeech 2019*, pages 1841–1845.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147, Los Angeles. Association for Computational Linguistics.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic. Association for Computational Linguistics.

Thomas Schatz. 2016. *ABX-discriminability measures and applications*. Ph.D. thesis, Université Paris 6 (UPMC).

Leslie N Smith. 2017. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472. IEEE.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2020. Transformer VQ-VAE for Unsupervised Unit Discovery and Speech Synthesis: ZeroSpeech 2020 Challenge. *arXiv:2005.11676 [cs, eess]*. ArXiv: 2005.11676.

Andros Tjandra, Berrak Sisman, Mingyang Zhang, Sakriani Sakti, Haizhou Li, and Satoshi Nakamura. 2019. VQVAE Unsupervised Unit Discovery and Multi-scale Code2Spec Inverter for Zerospeech Challenge 2019. *arXiv:1905.11449 [cs, eess]*. ArXiv: 1905.11449.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6306–6315.

Benjamin van Niekerk, Leanne Nortje, and Herman Kamper. 2020. Vector-quantized neural networks for acoustic unit discovery in the zerospeech 2020 challenge. *arXiv preprint arXiv:2005.09409*.

Maarten Versteegh, Roland Thiolliere, Thomas Schatz, Xuan Nga Cao, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux. 2015. The zero resource speech challenge 2015. In *Sixteenth annual conference of the international speech communication association*.

A Supplementary material

A.1 Loss of a logistic diagnostic classifier

The most general measure of the amount of information about the value of a random variable Y obtained through the observation of the value of random variable X is mutual information $I(Y;X)$. Here we relate the loss of a logistic diagnostic classifier predicting Y from X in the special case where both Y and X are discrete, with image $\mathcal{Y}$ and $\mathcal{X}$ respectively.

We can construct a logistic classifier which outputs the empirical probability $\hat{P}(Y=y|X=x)$ with $y \in \mathcal{Y}$ and $x \in \mathcal{X}$, by using a one-hot encoding of the categorical predictor variable X as $\mathbf{x}$ and by setting the classifier coefficients $\mathbf{W}$ as

$$W_{y,x} = \ln \hat{P}(Y=y|X=x). \quad (4)$$

The softmax of the logistic classifier with these coefficients simplifies to the empirical estimates of conditional probabilities of Y:

$$p_{y|\mathbf{x}} = \frac{\exp(\ln \hat{P}(Y=y|X=x))}{\sum_{z \in \mathcal{Y}} \exp(\ln \hat{P}(Y=z|X=x))} \quad (5)$$
$$= \hat{P}(Y=y|X=x). \quad (6)$$

The cross entropy of the predictions of this classifer is then:

$$J(\mathbf{w}) = -\frac{1}{N} \sum_{n=1}^{N} \ln \hat{P}(Y=y_n|X=x_n) \quad (7)$$

where y_n and x_n are the values taken by the random variables Y and X for the nth example. The loss is equivalent to the empirical estimate of the conditional entropy $H(Y|X)$, and related to the mutual information between $I(Y;X)$ via:

$$I(Y;X) = H(Y) - H(Y|X). \quad (8)$$

To the extent that the scores of the logistic diagnostic classifier and normalized mutual information applied to the same data are not perfectly correlated, this would be due to the stochasticity of training, regularization, as well as the use of accuracy rather than cross entropy to measure performance.

A.2 Overall performance of the visually supervised model

Since the visually-supervised model is trained and optimized for matching images and their corresponding spoken captions, we measure the recall@10 of retrieving the correct image for a given
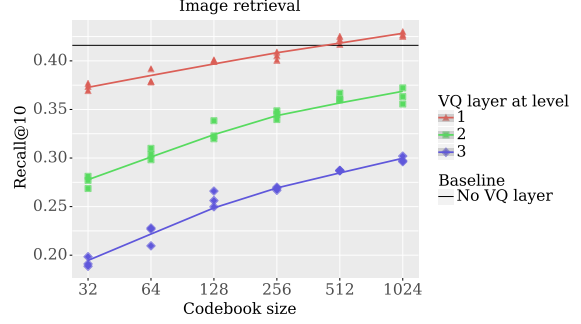

Figure 2: Recall@10 as a function of codebook size and level of placement.

Model	Input	Size	Skew	Kurtosis
SS	complete	32	0.14	0.21
SS	triplet	32	-0.67	0.55
SS	complete	1024	-0.20	0.29
SS	triplet	1024	-2.79	8.99
VS	complete	32	0.04	0.35
VS	triplet	32	-1.66	3.20
VS	complete	1024	-1.40	4.71
VS	triplet	1024	-7.92	87.20

Table 4: Skew and excess kurtosis of edit distance distributions, for the self-supervised (SS) and visually-supervised level 1 (VS) models.

spoken utterance as a function of the size of the learned codebook and the placement of the VQ layer. Figure 2 shows these results. We observed the following patterns:

- Most models perform worse than a model without the VQ layer, with the exception of the models with the VQ layer at level 1 and a codebook of size 512 or 1024.

- Performance on the image retrieval task is negatively correlated with the level of placement of the VQ layer.

- VQ layers with larger codebooks perform better.

A.3 Edit distance distribution

Table 4 presents skew and excess kurtosis of edit distance distributions for both target models, using codebooks of size 32 and 1024 and trained on long and short segments, confirming the hypothesis that the combination of short segments and large codebooks leads to skewed distributions.

A.4 Speaker identification

Figure 3 shows accuracy of diagnostic classifiers trained on code sequences encoded as vectors of code frequencies. For visually-supervised models, speaker identity is represented to some degree in untrained models, for codebooks of all sizes and at all levels, and most strongly for large codebook sizes at level 1. After training, we observe differentiated patterns for VQ layers at level 1 versus 2 and 3. While speaker identity is emphasized in codebooks at level 1 compared to the untrained models, it is weakened for subsequent layers, to the point of being effectively removed at level 3. These results indicate that VQ layers at level 1 represent speaker-dependent information, possibly encoding acoustic rather than phonemic information. The self-supervised models show results similar to the visually-supervised models with the VQ layer at level 1 when trained, but capture nearly no speaker information before training.
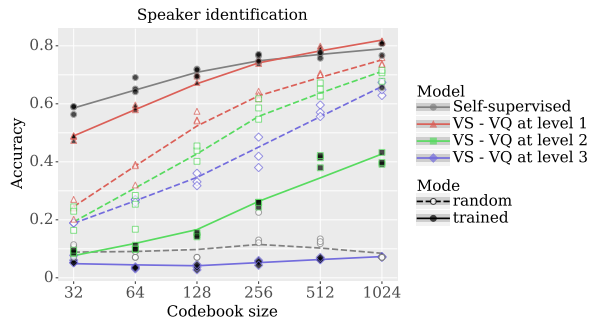

Figure 3: Accuracy of speaker classification on codebooks for the self-supervised and visually-supervised models. Trained models are represented by black icons and solid lines. Dashed lines and white filled icons represent randomly initialized models.