LEARNING HIERARCHICAL DISCRETE LINGUISTIC UNITS FROM VISUALLY-GROUNDED SPEECH

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we present a method for learning discrete linguistic units by incorporating vector quantization layers into neural models of visually grounded speech. We show that our method is capable of capturing both word-level and sub-word units, depending on how it is configured. What differentiates this paper from prior work on speech unit learning is the choice of training objective. Rather than using a reconstruction-based loss, we use a discriminative, multimodal grounding objective which forces the learned units to be useful for semantic image retrieval. We evaluate the sub-word units on the ZeroSpeech 2019 challenge, achieving a 27.3% reduction in ABX error rate over the top-performing submission, while keeping the bitrate approximately the same. We also present experiments demonstrating the noise robustness of these units. Finally, we show that a model with multiple quantizers can simultaneously learn phone-like detectors at a lower layer and word-like detectors at a higher layer. We show that these detectors are highly accurate, discovering 279 words with an F1 score of greater than 0.5.

1 INTRODUCTION

By 8 months of age, human infants learn to recognize not only the names of their caregivers and common objects, but also the contrast between the different vowels and consonants which comprise these words (Dupoux, 2018). Nearly all toddlers learn to carry a conversation long before they can read and write. Humans learn to model the discrete, hierarchical, and compositional nature of their native language not from written text, but from speech audio - a continuous, time-varying waveform which is the product not only of the underlying words which were spoken, but also the physical properties of the speaker's vocal tract, the speaker's health and emotional state, and the noise and reverberation present in the environment. The question of how such a complex symbolic system is inferred from continuous and noisy sensory input data is of interest not only to the cognitive science community, but also to machine learning researchers who aim to reproduce this ability with computers. A more comprehensive understanding of human language acquisition has practical significance in real-world applications, such as automatic speech recognition (ASR) and natural language understanding (NLU) systems. In the past several decades, enormous progress has been made in speech recognition research, and nowadays ASR systems are able to achieve human-level accuracy in many domains (Chiu et al., 2018). Unfortunately, the techniques that have been developed to achieve these levels of performance are extremely data-hungry, requiring many thousands of hours of speech audio recordings for training. Since supervised machine learning algorithms form the basis of ASR training, the data also needs to be annotated by expert humans. Due to the immense cost of collecting and annotating speech data, ASR technology currently exists for approximately 120 (Google, 2019) out of the nearly 7,000 (Lewis et al., 2016) human languages spoken worldwide. It is highly unlikely that purely supervised machine learning techniques will be able to scale to include all human languages, necessitating the development of alternative methods by researchers which are able to function with far fewer annotations, or even no annotations at all. Because human beings provide an existence proof of language acquisition from speech completely without language supervision, it is plausible that this ability could be replicated by a machine learning algorithm.

In this paper, we present a method for discovering *discrete* and *hierarchical* representations of speech units both at the sub-word level and the word level. The key difference between our work and previously proposed linguistic unit discovery methods is the fact that we employ a discriminative, visual-semantic grounding objective rather than a signal reconstruction objective. This forces our

models to learn representations which capture semantic information at the output layer of the network. Because semantics are predominantly carried by words, and words are composed of sub-word units (such as phones and syllables), the visual grounding objective indirectly forces the model to learn speaker- and noise-invariant representations of speech units. By incorporating trainable quantization layers into our networks, we are able to capture these units in discrete inventories. Whether these units correspond to word-like or sub-word units depends on where the quantization layers are inserted, and how they are trained.

2 Related Work

Prior work on unsupervised modeling of the speech signal has generally focused on learning representations which either disentangle or isolate the latent factors that are of interest for downstream tasks. In most cases the primary latent factor of interest is the phonetic or lexical identity of a given segment of speech, but other factors, such as the identity of the speaker, are sometimes of interest as well. Because the factors of interest are often inherently discrete (e.g. words and phones), many of the proposed approaches attempt to perform segmentation and clustering of the surface features in one way or another. One family of techniques is based upon Segmental Dynamic Time Warping (S-DTW) (Park & Glass, 2005; 2008; Jansen et al., 2010; Jansen & Durme, 2011), which uses a self-comparison algorithm to identify relatively long duration (on the order of a second) patterns which frequently reoccur in a speech corpus; these patterns tend to capture words or short phrases. A different line of work employs probabilistic graphical models to jointly segment and cluster the speech signal (Varadarajan et al., 2008; Zhang & Glass, 2009; Gish et al., 2009; Lee & Glass, 2012; Siu et al., 2014; Lee et al., 2015; Ondel et al., 2016; Kamper et al., 2016; 2017a). With an appropriately designed model, it is possible to learn multiple, hierarchical categories of speech units. However, in order to enable efficient inference, the conditional distributions of these models tend to be simple and therefore have limited modeling power.

Deep neural network models have been successfully used to learn powerful speech representations using weakly or unsupervised objectives (Thiolliere et al., 2015; Kamper et al., 2015; Hsu et al., 2017a;b; Hsu & Glass, 2018; Holzenberger et al., 2018; Milde & Biemann, 2018; van den Oord et al., 2018; Chung et al., 2019; Pascual et al., 2019). These representations have predominantly been continuous in nature, as discrete latent variables are not trivially compatible with backpropagation. To obtain discrete representations, a post-hoc clustering step can be applied to the continuous representations (Kamper et al., 2017b; Feng et al., 2019). More recently, several papers have proposed ways of directly incorporating discrete variables into neural network models, including using Gumbel-Softmax (Eloff et al., 2019b) or straight-through estimators (van den Oord et al., 2017; Chorowski et al., 2019; Razavi et al., 2019).

A different method for learning meaningful representations of speech is via a multimodal grounding objective, which encourages the learning of speech representations that are predictive of the contextual information contained in a separate but accompanying modality, such as vision. Visual grounding of speech is a form of self-supervised learning (Virginia de Sa, 1994), which is powerful in part because it offers a way of training models with a discriminative objective that does not depend on traditional transcriptions or annotations. The first work in this direction relied on phone strings to represent the speech (Roy & Pentland, 2002; Roy, 2003), but more recently this learning has been shown to be possible directly on the speech signal (Synnaeve et al., 2014; Harwath & Glass, 2015; Harwath et al., 2016). Subsequent work on visually-grounded models of speech has investigated improvements and alternatives to the modeling or training algorithms (Leidal et al., 2017; Kamper et al., 2017c; Havard et al., 2019a; Merkx et al., 2019; Chrupała et al., 2017; Scharenborg et al., 2018; Kamper et al., 2019b;a; Surís et al., 2019; Ilharco et al., 2019; Eloff et al., 2019a), application to multilingual settings (Harwath et al., 2018a; Kamper & Roth, 2017; Azuh et al., 2019; Havard et al., 2019a), analysis of the linguistic abstractions, such as words and phones, which are learned by the models (Harwath & Glass, 2017; Harwath et al., 2018b; Drexler & Glass, 2017; Alishahi et al., 2017; Harwath et al., 2019; Harwath & Glass, 2019; Havard et al., 2019b), and the impact of jointly training with textual input (Holzenberger et al., 2019; Chrupała, 2019; Pasad et al., 2019). Representations learned by models of visually grounded speech are also well-suited for transfer learning to supervised tasks, being highly robust to noise and domain shift (Hsu et al., 2019).

3 DATA AND MODELS

3.1 DATASET

For training our models, we utilize the MIT Places 205 dataset (Zhou et al., 2014) and their accompanying spoken audio captions (Harwath et al., 2016; 2018b). The caption dataset contains approximately 400,000 spoken audio captions, each of which describes a different Places image. These captions are free-form spontaneous speech, collected from over 2,500 different speakers and covering a 40,000 word vocabulary. The average caption duration is approximately 10 seconds, and each caption contains on average 20 words. For vetting our models during training, we use a held-out validation set of 1,000 image-caption pairs.

3.2 NEURAL MODELS OF VISUALLY-GROUNDED SPEECH

We base our model upon the Residual Deep Audio-Visual Embedding network (ResDAVEnet) architecture (Harwath et al., 2019), which contains two branches of fully convolutional networks, one for images and the other for audio. Each branch encodes samples of the corresponding modality into a d-dimensional space, regardless of the original dimensionality of the samples. This is achieved by applying global spatial mean pooling and global temporal mean pooling to the image branch output and the audio branch output, respectively. The image branch is adapted from ResNet50 (He et al., 2016), where the final softmax layer and the preceding fully-connected layers are removed, replaced with a 1x1 linear convolutional layer in order to project the feature map to the desired dimension. To model the audio inputs, a 17-layer fully convolutional network with residual connections is used. The input is a log Mel-frequency spectrogram with 40 frequency bins and 25 ms-wide, Hamming-windowed frames with a shift of 10 ms. The first layer of this network is a 1-D convolution that spans the entire frequency axis of the spectrogram, while the remaining 16 convolutional layers are 1-D across the time axis. These 16 layers are divided into four residual blocks of 4 layers each, and downsampling between these blocks is accomplished by applying the first convolution of each block with a stride of 2. For full details of the model, refer to Harwath et al. (2019).

3.3 LEARNING HIERARCHICAL DISCRETE UNITS WITH VECTOR QUANTIZING LAYERS

Previous analyses reveal that ResDAVEnet-like models learn linguistic abstractions at different levels, including words (Harwath & Glass, 2017) and robust phonetic features (Harwath & Glass, 2019; Hsu et al., 2019). To explicitly learn hierarchical discrete linguistic units within this framework, we propose to incorporate multiple vector quantization (VQ) layers (van den Oord et al., 2017) into the ResDAVEnet audio branch; we refer to this new architecture as ResDAVEnet-VQ.

VQ layers can be understood as a type of bottleneck, which constrain the amount of information that can flow through. While these layers have been used to learn discrete sub-word units (van den Oord et al., 2017; Chorowski et al., 2019; Razavi et al., 2019), previous work injects VQ layers into autoencoders that are trained with a reconstruction loss. As a result, the embedding dimension of each code and the number of codes need to be carefully tuned (Liu et al., 2019). When the embedding dimension is too low or the codebook size too small, the model does not have enough expressive power to capture linguistic variability. When it is too large, the model starts to encode non-linguistic information in order to improve reconstruction. In contrast, the learning signal of ResDAVEnet-VQ is provided by the visual-semantic grounding objective. Rather than encoding as much information about input as possible, the learned codes in ResDAVEnet-VQ only need to capture semantic information. Since semantics in speech are predominantly transmitted by words, and words are composed of sub-word units like phones, the grounding objective places pressure on the model to robustly infer both from speech. Since words and phones are inherently discrete symbols, representing them with learned discrete units may not even hurt the grounding performance.

Figure 1 illustrates the proposed ResDAVEnet-VQ model. We add a quantization layer after each of the first two residual blocks of the ResDAVEnet-VQ model, denoted as VQ2 and VQ3, respectively, with the intention that they should capture discrete word-like and sub-word-like units. A VQ layer is defined as $\boldsymbol{E} \in \mathbb{R}^{K \times D}$, where K represents the codebook size, and D represents the output dimensionality of the input features to the codebook. Denoting the tth temporal frame of the input to the quantization layer as $\boldsymbol{x}_t$, quantization is performed according to $\boldsymbol{q}_t = \boldsymbol{E}_{k,:}$, where $k = \arg\min_j ||\boldsymbol{x}_t - \boldsymbol{E}_{j,:}||_2$ The quantized output is then fed as input to the subsequent residual block. As

in van den Oord et al. (2017), we use the straight-through estimator (Bengio et al., 2013) to compute the gradient passed from $\boldsymbol{q}_t$ to $\boldsymbol{x}_t$. We use the exponential moving average (EMA) codebook updates proposed by van den Oord et al. (2017).

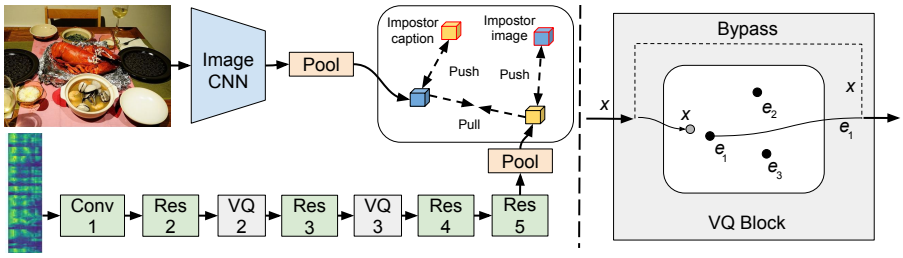

Figure 1: Diagram of the ResDAVEnet-VQ model. On the left, we show the placement of the vector quantization blocks in the audio branch. The right half of the figure depicts the quantization mechanism of each VQ block, as well as the bypass path when the block is disabled. For model details not relating to the VQ blocks, refer to Harwath et al. (2019).

3.4 Codebook Learning Schedules

We include multiple VQ layers in the ResDAVEnet-VQ model, each of which can be independently enabled or bypassed without changing the rest of the architecture configuration. A model can be warm-started by copying the weights from another trained model that has fewer VQ layers enabled, and randomly initializing the codebook of the newly activated VQ layer(s). This gives rise to the questions of how many quantizers should be used and in what order they should be enabled. It is unclear whether models with same VQ layers activated would learn the same representation at each layer regardless of the training curriculum. Let A_m denote a subset of all VQ layers, and $A_{m-1} \subset A_m$. We use "$A_1 \rightarrow ... \rightarrow A_M$" to denote a model that is obtained by sequentially training models "$A_1 \rightarrow ... \rightarrow A_m$" initialized from "$A_1 \rightarrow ... \rightarrow A_{m-1}$", where the model A_1 is initialized from scratch, and the final model would have VQ layers in A_M activated. For instance, a model initialized from scratch with no VQ layers enabled is denoted as "$\varnothing$", and a model initialized with that and with both layers enabled is denoted as "$\varnothing \rightarrow \{2, 3\}$".

3.5 Training with the Triplet Loss

We train our models using the same loss function as Harwath et al. (2019). This loss function blends two triplet loss terms (Weinberger & Saul, 2009), one based on random sampling of negative examples, and the other based on semi-hard negative mining (Jansen et al., 2018), in order to find more challenging negative samples. Specifically, let the sets of output embedding vectors for a minibatch of B audio/image training pairs respectively be $\mathbb{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_B\}$ and $\mathbb{I} = \{\boldsymbol{i}_1, \ldots, \boldsymbol{i}_B\}$. To compute the randomly-sampled triplet loss term, we select impostor examples for the jth input according to $\bar{\boldsymbol{a}}_j \sim \text{UniformCategorical}(\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_B\} \setminus \boldsymbol{a}_j)$ and $\bar{\boldsymbol{i}}_j \sim \text{UniformCategorical}(\{\boldsymbol{i}_1, \ldots, \boldsymbol{i}_B\} \setminus \boldsymbol{i}_j)$. The randomly-sampled triplet loss is then computed as:

$$\mathcal{L}_s = \sum_{j=1}^{B} \Big(\max(0, \boldsymbol{i}_j^T \bar{\boldsymbol{a}}_j - \boldsymbol{i}_j^T \boldsymbol{a}_j + 1) + \max(0, \bar{\boldsymbol{i}}_j^T \boldsymbol{a}_j - \boldsymbol{i}_j^T \boldsymbol{a}_j + 1) \Big) \tag{1}$$

For the semi-hard negative triplet loss, we first define the sets of impostor candidates for the jth example as $\hat{\mathbb{A}}_j = \{\boldsymbol{a} \in \mathbb{A} | \boldsymbol{i}_j^T \boldsymbol{a} < \boldsymbol{i}_j^T \boldsymbol{a}_j\}$ and $\hat{\mathbb{I}}_j = \{\boldsymbol{i} \in \mathbb{I} | \boldsymbol{i}^T \boldsymbol{a}_j < \boldsymbol{i}_j^T \boldsymbol{a}_j\}$. The semi-hard negative loss is then computed as:

$$\mathcal{L}_h = \sum_{j=1}^{B} \Big(\max(0, \max_{\hat{\boldsymbol{a}} \in \hat{\mathbb{A}}_j} (\boldsymbol{i}_j^T \hat{\boldsymbol{a}}) - \boldsymbol{i}_j^T \boldsymbol{a}_j + 1) + \max(0, \max_{\hat{\boldsymbol{i}} \in \hat{\mathbb{I}}_j} (\hat{\boldsymbol{i}}^T \boldsymbol{a}_j) - \boldsymbol{i}_j^T \boldsymbol{a}_j + 1) \Big) \tag{2}$$

Finally, the overall loss function is computed as the sum of the two above losses, $\mathcal{L} = \mathcal{L}_s + \mathcal{L}_h$.

3.6 IMPLEMENTATION DETAILS

All of our models were trained for 180 epochs using the Adam optimizer (Kingma & Ba, 2014) with a batch size of 80. We used an exponentially decaying learning rate schedule, with an initial value of 2e-4 that decayed by a factor of 0.95 every 3 epochs. Following van den Oord et al. (2017), we use an EMA decay factor of $\gamma = .99$ for training each VQ codebook. Our core experimental results all use a codebook size of 1024 vectors for all quantizers, but in the supplementary material we include experiments with smaller and larger codebooks. Following Chorowski et al. (2019), the jitter probability hyperparameter for each quantization layer was fixed at 0.12. While we do not apply data augmentation to the input spectrograms, during training we perform standard data augmentation techniques to the images. We resize each raw image so that its smallest dimension is 256 pixels, and then we apply an Inception-style random crop which is resized to 224 pixels square. During training, we also flip each image horizontally with a probability of 0.5. During evaluation, the center 224 pixel square crop is always taken from the image. Finally, the RGB pixel values are mean and variance normalized. We trained each model on the Places audio caption train split, and computed the image and caption recall at 10 (R@10) scores on the validation split of the Places audio captions after each training epoch. The model snapshot that achieved the highest average R@10 score on the validation set from each training is used for all evaluation. To extract embeddings and units from our models, we simply perform a forward pass through the speech branch of the ResDAVEnet-VQ network and retain the outputs from the target layer at a uniform frame-rate. The frame-rate is determined by the downsampling factor at the target layer relative to the input. For non-quantized layers, these outputs will be continuous embeddings. For quantized layers, these will be quantized embedding retrieved from the assigned entry in the codebook.

4 EXPERIMENTS

4.1 SUB-WORD UNIT LEARNING ON THE ZEROSPEECH 2019 ABX TASK

Evaluation metrics Learning unsupervised speech representations that are indicative of phonetic content is of high interest to the speech community, and recently has been the focus of the ZeroSpeech Challenge (Versteegh et al., 2015; Dunbar et al., 2017; 2019). One of the core evaluations is the model-free, minimal-pair ABX task (Schatz et al., 2013), which aims to benchmark representations in terms of their discriminability between different sub-word speech units. In this task, three speech waveform segments denoted by A, B, and X are presented to a model. A and B are constrained to be a triphone minimal pair; that is, both segments capture three phones, but differ only in the identity of their center phone. The third segment, X is chosen to contain the same underlying triphone sequence as A. Supposing $f(\cdot)$ is a function to be evaluated that maps a waveform segment to a feature segment, the ABX error rate under a given similarity metric $S(\cdot, \cdot)$ is defined as the fraction of ABX triples in which $S(f(A), f(X)) > S(f(B), f(X))$. An ABX error rate of 50% indicates random assignment, while an ABX of 0% reflects perfect phone discriminability.

In the ZeroSpeech challenge, dynamic time warping (DTW) is used to measure similarity between a pair of feature sequences for ABX evaluations, with several options available for computing the distance between a pair of feature frames. The ZeroSpeech 2019 challenge in particular emphasizes on discovering an inventory of discrete sub-word units, rather than continuous representations. Therefore, in addition to an ABX error rate, a bitrate is also computed for each model which reflects the amount of information carried by the learned units. A lower bitrate can be achieved by having a more compact inventory of learned units or having a smaller number of codes per second. The full details of the evaluation can be found in Dunbar et al. (2019). To be clear, all of our ResDAVEnet-VQ models were not trained on the ZeroSpeech training data, but instead on the Places audio captions, thus there is a domain mismatch between training and testing these models.

In addition to the frame-based bitrate and ABX scores computed by the ZeroSpeech 2019 evaluation toolkit, we implement our own extensions to these metrics. Because it is common for successive frames to be assigned to the same codebook entry and phonetic information is not encoded at a fixed frame rate, lossless run length encoding (RLE) can be a more reasonable measure of the bitrate of a frame-based model. RLE does not change the ABX score since it can be trivially inverted, but it does change the bitrate. For computing the RLE bitrate, we modify the bitrate calculation specified in Dunbar et al. (2019) so that a unique symbol is defined as the tuple (unit, length) where length is

Table 1: Comparison of R@10, ABX scores, and bit-rates between different configurations and baseline models trained on ZeroSpeech 2019 data or Places Audio Caption. All quantizers reflected in this table used a codebook size of 1,024 vectors. We do not compute RLE or segment scores for the FHVAE-DPGMM model, since we did not re-implement that model.

Model ID	Layer	R@10	Frame-Based			Segment-Based	
			ABX	Bitrate	RLE Bitrate	ABX	Bitrate
FHVAE-DPGMM (ZS)	-	N/A	21.67	413.23	-	-	-
WaveNet-VQ (ZS)	-	N/A	19.98	151.55	136.74	20.48	126.17
WaveNet-VQ (PA)	-	N/A	24.87	149.00	136.27	25.23	126.22
"$\{2\}$"	VQ2	.753	12.33	433.30	361.09	12.78	332.86
"$\varnothing \rightarrow \{2\}$"	VQ2	.760	**11.79**	390.61	317.66	12.66	289.11
"$\{3\}$"	VQ3	.734	38.21	213.92	129.65	38.68	108.84
"$\varnothing \rightarrow \{3\}$"	VQ3	.794	**15.04**	182.93	140.04	16.53	121.26
"$\{2,3\}$"	VQ2	.667	25.62	408.75	258.37	26.32	217.58
	VQ3		32.23	218.76	156.69	32.49	136.90
"$\varnothing \rightarrow \{2,3\}$"	VQ2	.787	13.15	405.43	334.39	13.30	303.03
	VQ3		14.95	199.91	172.05	15.60	159.07
"$\{2\} \rightarrow \{2,3\}$"	VQ2	.764	**12.51**	415.13	341.85	13.06	311.82
	VQ3		**14.52**	167.84	136.11	15.68	121.17
"$\{3\} \rightarrow \{2,3\}$"	VQ2	.760	13.55	421.23	271.91	14.38	232.87
	VQ3		33.70	208.63	117.37	33.58	98.29

the number of frames assigned to a given unit with in a segment. We also consider segment-based ABX and bitrate, which is similar to the RLE metrics except in this case we outright discard the frame length information. This typically results in an even greater reduction in bitrate, but also an accompanying deterioration in ABX score.

Baseline models In Table 1, we compare our results to those derived from two of the top-performing submissions to the ZeroSpeech 2019 challenge: a re-implementation of WaveNet-VQ (Chorowski et al., 2019) provided by Cho et al. (2019) and FHVAE-DPGMM (Feng et al., 2019). Using the code accompanied with the WaveNet-VQ submission, we were able to train their model on the set of 400,000 Places audio captions to make a fairer comparison with our ResDAVEnet-VQ models in terms of the amount of speech data used. In addition, when trying to reproduce the reported WaveNet-VQ results, we obtain better performance than previously reported by training for more steps. Table 1 shows that WaveNet-VQ achieves similar bitrates regardless of the training data. However, ABX deteriorates from 19.98 to 24.87, implying the model cannot utilize data of a larger scale but out-of-domain relative to the test set. A similar degradation when testing on out-of-domain data with FHVAE models was observed in Hsu et al. (2019). We did not re-train the model submitted by Feng et al. (2019), and instead compare against the scores reported in Dunbar et al. (2019).

ABX discrimination without using quantization Our first experiment investigates exactly which layer in the ResDAVEnet-VQ model is most suited for ABX phone discrimination, and would thus make a good candidate for learning of quantized sub-word units. The leftmost plot in Figure 2 shows that layers 2 and 3 of a ResDAVEnet-VQ model without any quantization enabled perform the best in terms of ABX error rate on the ZeroSpeech 2019 English test set; the exact numbers for this model are displayed in the caption of Figure 2. Because layers 2 and 3 achieve the lowest ABX error rates without quantization, we focus our attention on the impact of quantization there.

Quantizing one layer When quantizing only one layer, we examine quantization of layer 2 vs. layer 3, and using cold-start training vs. warm-start initialization from model "$\varnothing$". The ABX and bitrate results for these models, as well as the R@10 scores on the Places validation set, are shown in Table 1. In all cases, quantization applied at the output of layer 2 achieves a better ABX score than quantization at layer 3, but VQ3 achieves a better bitrate. Quantization barely impacts the performance of layer 2, whose ABX score very slightly rises from 11.35 to 11.79. Warm-start initialization is beneficial to R@10 and ABX score in both cases, but we notice an intriguing anomaly when applying cold-start quantization to layer 3: the ABX score deteriorates significantly, rising from 10.86

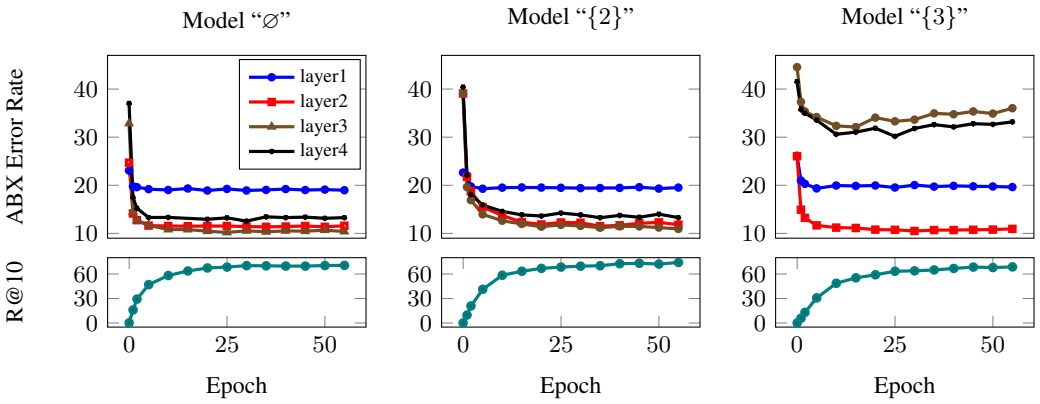

Figure 2: R@10 and ABX tracked at various training epochs. The "∅" model achieves a final R@10 of .735, with ABX scores of 19.77, 11.35, 10.86, and 14.05 for the conv1, res2, res3, and res4 layers.

in the case of the non-quantized model to 38.21. This indicates that while VQ2 is capable of learning a finite inventory of units that are highly predictive of phonetic identity from either a warm-start or cold-start initialization, cold-start training of VQ3 results in very little phonetic information captured by the quantizer. Interestingly, this model is still learning to infer visual semantics from the speech signal, as evidenced by a high R@10 score; we later show in Section 4.2 that the reason for this anomaly is because cold-start training of VQ3 results in the learning of word detectors. In all cases except for model "$\{3\}$", we note that the ABX scores achieved by our models are significantly better than the baselines. Our best model in terms of ABX ("$\varnothing \rightarrow \{2\}$") achieves a 41.0% reduction in ABX over the WaveNet-VQ baseline, at a cost of a 132.3% increase in RLE bitrate; however, model "$\varnothing \rightarrow \{3\}$" achieves a 24.7% reduction in ABX error rate with only a 2.4% increase in RLE bitrate. These results do not constitute a fair comparison, however, because the WaveNet-VQ and ResDAVEnet-VQ models were trained on different datasets; when training the WaveNet-VQ model on the same set of audio captions used to train ResDAVEnet-VQ (but without the accompanying images, since WaveNet-VQ is not a multimodal model), the ABX error rate increases to 24.87%, tipping the results even more in favor of the ResDAVEnet-VQ models.

Quantizing two layers Quantizing multiple layers at once offers the possibility of learning a hierarchy of units. Thus, we aim to capture phonetic information in a lower layer quantizer and word-level information at a higher layer quantizer. Cold-start training of two quantizers ("$\{2, 3\}$") results in a significant drop in ABX performance for both VQ2 and VQ3, but also a drop in R@10 on the Places validation set. We see much better results in terms of R@10 and ABX for the remaining 3 models which were initialized from the "∅" model or a model with only one quantizer enabled; for example, model "$\{3\} \rightarrow \{2, 3\}$" achieves an ABX of 14.52 with an RLE bitrate of 136.11, representing a 27.3% ABX improvement over the best baseline while keeping the bitrate approximately the same. We see in model "$\{3\} \rightarrow \{2, 3\}$" that the same phenomenon observed with model "$\{3\}$" persists: VQ3 achieves relatively poor ABX, despite a high overall R@10 and strong ABX with VQ2 at 13.55%. We confirm in Section 4.2 that the VQ3 layer of model "$\{3\} \rightarrow \{2, 3\}$" does indeed capture word-level information, indicating that this model has successfully localized phonetic unit identity in the second layer and lexical unit identity in the third layer. Overall, our results suggest that when learning hierarchical quantized representations with a ResDAVEnet-VQ model, the nature of the representations learned is highly dependent on the training curriculum.

Training and testing on noisy data In Hsu et al. (2019), it was shown that representations learned by a ResDAVEnet model were far more robust to train/test domain mismatch in terms of background noise, channel characteristics, and speaker identity than standard spectral features when training a supervised speech recognizer. Here, we examine whether this robustness is also exemplified by the quantized versions of this model. We construct three additional test sets using the ZeroSpeech 2019 English testing data by adding noise sampled from the AudioSet (Jansen et al., 2018) dataset. For each ZeroSpeech testing waveform, we randomly sampled an AudioSet waveform of the same duration and performed linear mixing with a signal-to-noise ratio (SNR) selected randomly within

Model	Layer	Clean		20-30 dB		10-20 dB		0-10 dB	
		ABX	R-B	ABX	R-B	ABX	R-B	ABX	R-B
WaveNet-VQ (ZS)	N/A	19.98	136.74	21.22	141.07	27.51	144.28	42.55	126.96
WaveNet-VQ (PA)	N/A	24.87	136.27	27.18	137.70	33.29	132.34	42.67	110.50
"$\varnothing$"	Res2	11.35	N/A	11.63	N/A	13.17	N/A	19.44	N/A
"$\varnothing$"	Res3	10.86	N/A	11.16	N/A	12.96	N/A	19.43	N/A
"$\varnothing \rightarrow \{2\}$"	VQ2	11.79	317.66	12.15	325.40	14.62	332.21	23.96	327.15
"$\{2\} \rightarrow \{2,3\}$"	VQ2	12.51	341.85	12.56	350.28	14.82	362.73	25.02	330.54
"$\{2\} \rightarrow \{2,3\}$"	VQ3	14.52	136.11	14.73	137.68	17.44	143.14	27.68	133.13
"$\{3\} \rightarrow \{2,3\}$"	VQ2	13.55	271.91	13.65	272.46	15.69	267.70	24.06	244.52
"$\{3\} \rightarrow \{2,3\}$"	VQ3	33.70	117.37	32.56	118.22	34.65	115.40	39.82	102.48
"$\varnothing$" (n)	Res2	13.32	N/A	12.30	N/A	12.97	N/A	16.91	N/A
"$\varnothing$" (n)	Res3	11.85	N/A	11.90	N/A	12.44	N/A	16.09	N/A
"$\varnothing \rightarrow \{2\}$" (n)	VQ2	12.64	342.53	12.20	348.57	13.34	359.43	18.82	373.60
"$\{2\} \rightarrow \{2,3\}$" (n)	VQ2	13.42	365.89	13.71	359.14	14.57	370.67	18.78	392.10
"$\{2\} \rightarrow \{2,3\}$" (n)	VQ3	14.39	179.19	14.92	180.36	15.38	182.27	19.58	188.32
"$\{3\} \rightarrow \{2,3\}$" (n)	VQ2	16.52	223.28	16.47	223.61	17.75	225.72	22.68	230.01
"$\{3\} \rightarrow \{2,3\}$" (n)	VQ3	26.21	187.31	25.88	187.92	26.34	188.49	31.26	191.28

Table 2: ABX scores and RLE bitrates for various SNRs on the noisy ZeroSpeech19 English test set. "R-B" stands for "RLE-Bitrate," and (n) denotes a model trained on the noisy Places Audio dataset. For the WaveNet-VQ models, (ZS) and (PA) respectively denote training on the ZeroSpeech 19 English training set, and the clean Places Audio dataset.

a specified range. We construct low, medium, and high noise testing sets, corresponding to SNRs of 20-30 dB, 10-20 dB, and 0-10 dB. We then perform the ABX discrimination task on these noisy waveforms, displaying the results in Table 2. We find that for all models, a worsening SNR results in a deterioration in ABX performance. However, the ResDAVEnet-VQ models prove to be far more noise robust than the Wavenet-VQ model; even in the high noise testing set, the best ResDAVEnet-VQ model achieves an ABX of 23.96%, while the WaveNet-VQ models degrade to nearly-random ABX scores of 42.55% and 42.67%.

Given that a ResDAVEnet-VQ model trained on the "clean" Places Audio captions is highly robust to additive noise on the ABX discrimination task, we investigated whether adding noise to the Places Audio captions themselves would result in an even higher degree of noise robustness. To that end, we followed a similar data augmentation approach to create a noisy version of the Places Audio captions, where the SNR of each caption was randomly chosen to sit within the range of 0-30 dB. The bottom half of Table 2 shows the results of training several ResDAVEnet-VQ models on the noisy Places Audio captions and testing on the clean and noisy ZeroSpeech ABX tasks. In general, we observe a degradation ABX score in the clean conditions, but with a significantly higher degree of noise robustness in the noisier conditions.

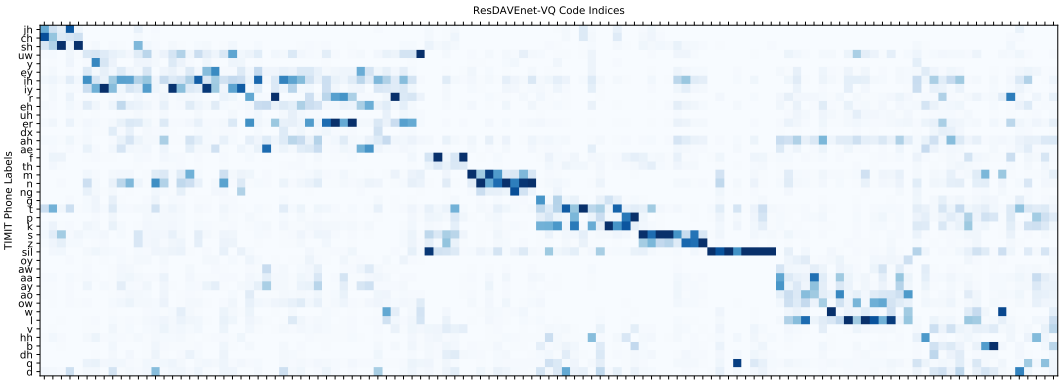

Figure 3: Conditional probability matrix displaying $P(phone|unit)$ using the "$\varnothing \rightarrow \{2\}$" model with a VQ2 codebook size of 128. For visualization, we saturate the color scaling at probability 0.5.

Visualization of learned units To better measure the correspondence between the VQ units and English phones, we compute corpus-level co-occurrence statistics (at the frame-level) across the TIMIT training set, excluding the `sa` dialect sentences. To facilitate visualization, we use the "$\varnothing \rightarrow \{2\}$" model with a codebook size of 128. We display the conditional probability matrix $P(phone|unit)$ in Figure 3, with the rows and columns ordered via spectral co-clustering with 10 clusters in order to group together phones that share similar sets of VQ codes. Visually, there is a strong mapping between TIMIT phone labels and ResDAVEnet-VQ codes. In some cases, redundant codes are used for the same phone label (this is especially the case for the silence label), and in other cases we see that phones belonging to the same manner class often tend to share codebook units. We can numerically quantify the mapping between the phone and unit labels with the normalized mutual information measure (NMI), which we found to be .378 in this case. We also include several spectrograms with their time-aligned unit sequences in the supplementary material.

Table 3: Performance of the VQ3 layer from the "$\{3\} \rightarrow \{2, 3\}$" model when codes are treated as word detectors. Codes are ranked by the highest F1 score among the retrieved words for a given code. Word hypotheses for a given code are ranked by the F1 score. P denotes precision, R recall, and occ the number of co-occurrences of the code and word in the data.

rank	code	Top Hypotheses					Second Hypotheses				
		word	F1	P	R	occ	word	F1	P	R	occ
1	58	baseball	90.09	89.45	90.75	3266	player	2.82	1.60	12.18	139
2	706	background	89.26	92.36	86.36	17740	backgrounds	0.71	0.36	68.75	66
3	88	classroom	88.78	87.39	90.21	1512	class	4.65	2.79	13.99	82
4	116	construction	88.16	90.19	86.22	2484	constructed	1.98	1.03	26.29	56
5	584	playground	87.84	85.68	90.12	930	play	2.70	1.81	5.33	51
					...						
198	234	alleyway	67.26	56.62	82.81	944	alley	37.25	30.06	48.98	863
199	1000	orange	67.25	89.09	54.00	2987	oranges	3.30	1.71	48.84	63
200	842	pink	67.19	78.12	58.94	2723	paint	5.05	3.07	14.26	140

Table 4: Performance of the VQ3 layer from the "$\{2\} \rightarrow \{2, 3\}$" model when codes are treated as word detectors.

rank	code	Top Hypotheses					Second Hypotheses				
		word	F1	P	R	occ	word	F1	P	R	occ
1	924	people	76.60	66.79	89.80	33662	computer	2.30	1.18	42.36	887
2	530	building	75.96	65.57	90.27	35229	buildings	23.44	13.44	91.70	7260
3	749	white	75.86	66.14	88.93	45097	one	5.22	3.15	15.33	3315
4	505	blue	58.47	46.53	78.67	21678	pool	9.13	4.97	56.87	2674
5	581	snow	58.30	42.59	92.38	9746	small	16.17	11.72	26.06	5617
					...						
198	329	standing	5.29	2.80	49.17	7769	woman	3.15	1.65	35.41	5704
199	362	white	5.27	3.05	19.49	9883	table	4.71	2.49	44.61	7351
200	256	trees	5.24	2.88	28.84	8963	see	4.48	2.45	26.12	6346

4.2 From Phones to Words: Learning a Hierarchy of Units

As shown in Table 1, all of the ResDAVEnet-VQ models which underwent cold-start training of VQ3 exhibited a similar phenomenon in which the ABX error rate of that layer was particularly high, despite the model performing well at the image-caption retrieval task. We hypothesized that this could be due to VQ3 learning to recognize higher level linguistic units, such as words. To examine this empirically, we inferred the VQ3 unit sequence for every audio caption in the Places Audio training set according to several different models. Using the estimated word-level transcriptions of the utterances (provided by the Google SpeechRecognition API), we computed precision, recall, and F1 scores for every unique (word, VQ3 code) pair for a given model and quantization layer. We then ranked the VQ codes in descending order according to their maximum F1 score for any word in the vocabulary. Table 3 shows a sampling of these statistics for model "$\{3\} \rightarrow \{2, 3\}$", Table 4

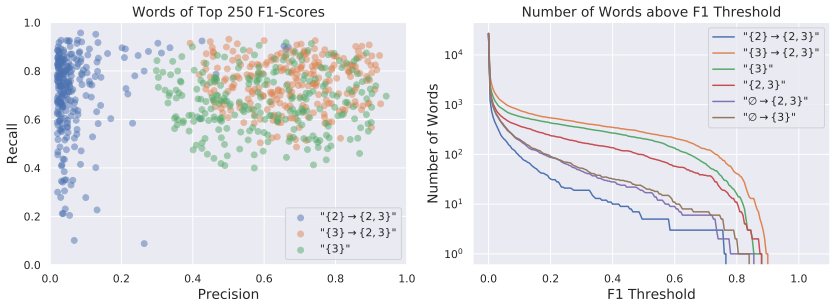

Figure 4: Visualization of the precision, recall, and F1 scores of individual VQ3 codes when treated as word detectors on the Places Audio captions.

for model "$\{2\} \rightarrow \{2, 3\}$". It should be emphasized that these models are exactly the same in all respects, except for the order in which their quantizers were trained.

We examine the overall performance of VQ3 as a word detector for these models in Figure 4. The right hand side of Figure 4 displays the number of VQ3 codes whose maximum F1 score is above a given threshold, while the left hand side shows the distribution of precision and recall scores for the top 250 words ranked by F1. This gives an approximate indication of how many VQ3 codes have learned to specialize as detectors for a specific word. We see that the VQ3 layer of model "$\{3\} \rightarrow \{2, 3\}$" learns 279 codebook entries with an F1 score above 0.5. In contrast, the VQ3 layer of model "$\{2\} \rightarrow \{2, 3\}$" learns only a handful of word-detecting codebook entries with an F1 of greater than 0.5. This experiment supports the notion that the reason for the poor ABX performances of the VQ3 layer in models "$\{3\}$" and "$\{3\} \rightarrow \{2, 3\}$" is in fact due to its specialization for detecting specific words, and that this specialization only emerges when the VQ3 layer is learned before the VQ2 layer. More extensive tables detailing the words learned by many more VQ3 codes for these models can be found in the supplementary material.

5 Conclusions

In this paper, we demonstrated that the neural vector quantization layers proposed by van den Oord et al. (2017) can be integrated into the visually-grounded speech models proposed by Harwath et al. (2019). This resulted in the ability of the speech model to directly represent speech units, such as phones and words, as discrete latent variables. We presented extensive experiments and analysis of these learned representations, demonstrating significant improvements in phone discrimination ability over the current state-of-the-art models for sub-word speech unit discovery. We demonstrated that these units are also far more robust to noise and domain shift than units derived from previously proposed models. These results supported the notion that semantic supervision via a discriminative, multimodal grounding objective has the potential to be more powerful than reconstruction-based objectives typically used in unsupervised speech models.

We also showed how multiple vector quantizers could be employed simultaneously within a single ResDAVEnet-VQ model, and that these quantizers could be made to specialize in learning a hierarchy of speech units: specifically, phones in the lower quantizer and words in the upper quantizer. Our analysis showed that hundreds of codebooks in the upper quantizer learned to perform as word detectors, and that these detectors were highly accurate. Our experiments also revealed that this behavior only emerged when VQ3 was trained before VQ2. These results suggest the importance of the learning curriculum, which should be more deeply investigated in future work. Future work should attempt to make explicit what kind of compositional rules are implicitly encoded by these models when mapping sequences of codes from the lower quantizer to word-level units in the upper quantizer; the automatic derivation of a sub-word unit inventory, vocabulary, and pronunciation lexicon could serve as the starting point for a fully unsupervised speech recognition system. Future work should also investigate whether layers above VQ3 could be made to learn even higher-level linguistic abstractions, such as grammar, syntax, and compositional reasoning.

REFERENCES

Afra Alishahi, Marie Barking, and Grzegorz Chrupała. Encoding of phonology in a recurrent neural model of grounded speech. In *Proc. ACL Conference on Natural Language Learning (CoNLL)*, 2017.

Emmanuel Azuh, David Harwath, and James Glass. Towards bilingual lexicon discovery from visually grounded speech audio. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019.

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

Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al. State-of-the-art speech recognition with sequence-to-sequence models. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

Suhee Cho, Yeonjung Hong, Yookyunk Shin, and Youngsun Cho. VQVAE with speaker adversarial training, 2019. URL `https://github.com/Suhee05/Zerospeech2019`.

Jan Chorowski, Ron J. Weiss, Samy Bengio, and Aäron van den Oord. Unsupervised speech representation learning using wavenet autoencoders. *IEEE Transactions on Audio, Speech and Language Processing*, 2019.

Grzegorz Chrupała. Symbolic inductive bias for visually grounded learning of spoken language. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. Representations of language in a model of visually grounded speech signal. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James R. Glass. An unsupervised autoregressive model for speech representation learning. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019.

Jennifer Drexler and James Glass. Analysis of audio-visual features for unsupervised speech recognition. In *Proc. Grounded Language Understanding Workshop*, 2017.

Ewan Dunbar, Xuan Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. The zero resource speech challenge 2017. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2017.

Ewan Dunbar, Robin Algayres, Julien Karadayi, Mathieu Bernard, Juan Benjumea, Xuan-Nga Cao, Lucie Miskic, Charlotte Dugrain, Lucas Ondel, Alan W. Black, Laurent Besacier, Sakriani Sakti, and Emmanuel Dupoux. The zero resource speech challenge 2019: TTS without T. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019.

Emmanuel Dupoux. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. In *Cognition*, 2018.

Ryan Eloff, Herman Engelbrecht, and Herman Kamper. Multimodal one-shot learning of speech and images. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019a.

Ryan Eloff, Andr Nortje, Benjamin van Niekerk, Avashna Govender, Leanne Nortje, Arnu Pretorius, Elan van Biljon, Ewald van der Westhuizen, Lisa van Staden, and Herman Kamper. Unsupervised acoustic unit discovery for speech synthesis using discrete latent-variable neural networks. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019b.

Siyuan Feng, Tan Lee, and Zhiyuan Peng. Combining adversarial training and disentangled speech representation for robust zero-resource subword modeling. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019.

Herb Gish, Man-Hung Siu, Arthur Chan, and William Belfield. Unsupervised training of an HMM-based speech recognizer for topic classification. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2009.

Google. Google cloud speech-to-text API. `https://cloud.google.com/speech-to-text/`, 2019. Accessed: 2019-09-16.

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

David Harwath and James Glass. Learning word-like units from joint audio-visual analysis. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

David Harwath and James Glass. Towards visually grounded sub-word speech unit discovery. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

David Harwath, Antonio Torralba, and James R. Glass. Unsupervised learning of spoken language with visual context. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2016.

David Harwath, Galen Chuang, and James Glass. Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018a.

David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *Proc. IEEE European Conference on Computer Vision (ECCV)*, 2018b.

David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. *International Journal of Computer Vision*, 2019.

William Havard, Jean-Pierre Chevrot, and Laurent Besacier. Models of visually grounded speech signal pay attention to nouns: a bilingual experiment on English and Japanese. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019a.

William N. Havard, Jean-Pierre Chevrot, and Laurent Besacier. Word recognition, competition, and activation in a model of visually grounded speech. In *Proc. ACL Conference on Natural Language Learning (CoNLL)*, 2019b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Nils Holzenberger, Mingxing Du, Julien Karadayi, Rachid Riad, and Emmanuel Dupoux. Learning word embeddings: Unsupervised methods for fixed-size representations of variable-length speech segments. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2018.

Nils Holzenberger, Shruti Palaskar, Pranava Madhyastha, Florian Metze, and Raman Arora. Learning from multiview correlations in open-domain videos. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

Wei-Ning Hsu and James Glass. Scalable factorized hierarchical variational autoencoder training. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2018.

Wei-Ning Hsu, Yu Zhang, and James Glass. Learning latent representations for speech generation and transformation. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2017a.

Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2017b.

Wei-Ning Hsu, David Harwath, and James Glass. Transfer learning from audio-visual grounding to speech recognition. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019.

Gabriel Ilharco, Yuan Zhang, and Jason Baldridge. Large-scale representation learning from visually grounded untranscribed speech. In *Proc. ACL Conference on Natural Language Learning (CoNLL)*, 2019.

Aren Jansen and Benjamin Van Durme. Efficient spoken term discovery using randomized algorithms. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

Aren Jansen, Kenneth Church, and Hynek Hermansky. Toward spoken term discovery at scale with zero resources. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2010.

Aren Jansen, Manoj Plakal, Ratheet Pandya, Daniel P.W. Ellis, Shawn Hershey, Jiayang Liu, R. Channing Moore, and Rif A. Saurous. Unsupervised learning of semantic audio representations. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

Herman Kamper and Michael Roth. Visually grounded cross-lingual keyword spotting in speech. In *Proc. of the Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, 2017.

Herman Kamper, Aren Jansen, and Sharon Goldwater. Fully unsupervised small-vocabulary speech recognition using a segmental Bayesian model. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2015.

Herman Kamper, Aren Jansen, and Sharon Goldwater. Unsupervised word segmentation and lexicon discovery using acoustic word embeddings. *IEEE Transactions on Audio, Speech and Language Processing*, 2016.

Herman Kamper, Aren Jansen, and Sharon Goldwater. A segmental framework for fully-unsupervised large-vocabulary speech recognition. *Computer Speech and Language*, 46(3):154–174, 2017a.

Herman Kamper, Karen Livescu, and Sharon Goldwater. An embedded segmental k-means model for unsupervised segmentation and clustering of speech. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2017b.

Herman Kamper, Shane Settle, Gregory Shakhnarovich, and Karen Livescu. Visually grounded learning of keyword prediction from untranscribed speech. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2017c.

Herman Kamper, Aristotelis Anastassiou, and Karen Livescu. Semantic query-by-example speech search using visual grounding. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019a.

Herman Kamper, Gregory Shakhnarovich, and Karen Livescu. Semantic speech retrieval with a visually grounded model of untranscribed speech. *IEEE Transactions on Audio, Speech and Language Processing*, 2019b.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations (ICLR)*, 2014.

Chia-Ying Lee and James Glass. A nonparametric Bayesian approach to acoustic model discovery. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2012.

Chia-Ying Lee, Timothy J. O'Donnell, and James Glass. Unsupervised lexicon discovery from acoustic input. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2015.

Kenneth Leidal, David Harwath, and James Glass. Learning modality-invariant representations for speech and images. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2017.

M. Paul Lewis, Gary F. Simon, and Charles D. Fennig. *Ethnologue: Languages of the World, Nineteenth edition*. SIL International. Online version: http://www.ethnologue.com, 2016.

Andy T Liu, Po-chun Hsu, and Hung-yi Lee. Unsupervised end-to-end learning of discrete linguistic units for voice conversion. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019.

Danny Merkx, Stefan L. Frank, and Mirjam Ernestus. Language learning using speech to image retrieval. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019.

Benjamin Milde and Chris Biemann. Unspeech: Unsupervised speech context embeddings. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2018.

Lucas Ondel, Lukás Burget, and Jan Černocký. Variational inference for acoustic unit discovery. In *Proc. of the Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, 2016.

Alex Park and James Glass. Towards unsupervised pattern discovery in speech. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2005.

Alex Park and James Glass. Unsupervised pattern discovery in speech. *IEEE Transactions on Audio, Speech and Language Processing*, 2008.

Ankita Pasad, Bowen Shi, Herman Kamper, and Karen Livescu. On the contributions of visual and textual supervision in low-resource semantic speech retrieval. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019.

Santiago Pascual, Mirco Ravanelli, Joan Serrà, Antonio Bonafonte, and Yoshua Bengio. Learning problem-agnostic speech representations from multiple self-supervised tasks. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *arXiv preprint arXiv:1906.00446*, 2019.

Deb Roy. Grounded spoken language acquisition: Experiments in word learning. *IEEE Transactions on Multimedia*, 5(2):197–209, 2003.

Deb Roy and Alex Pentland. Learning words from sights and sounds: a computational model. *Cognitive Science*, 26:113–146, 2002.

Odette Scharenborg, Laurent Besacier, Alan W. Black, Mark Hasegawa-Johnson, Florian Metze, Graham Neubig, Sebastian Stüker, Pierre Godard, Markus Müller, Lucas Ondel, Shruti Palaskar, Philip Arthur, Francesco Ciannella, Mingxing Du, Elin Larsen, Danny Merkx, Rachid Riad, Liming Wang, and Emmanuel Dupoux. Linguistic unit discovery from multi-modal inputs in unwritten languages: Summary of the "Speaking Rosetta" JSALT 2017 workshop. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

Thomas Schatz, Vijayaditya Peddinti, Francis Bach, Aren Jansen, Hynek Hermansky, and Emmanuel Dupoux. Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2013.

Man-Hung. Siu, Herb Gish, Arthur Chan, William Belfield, and Steve Lowe. Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery. *Computer Speech and Language*, 28(1):210–223, 2014.

Dídac Surís, Adrià Recasens, David Bau, David Harwath, James Glass, and Antonio Torralba. Learning words by drawing images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Gabriel Synnaeve, Maarten Versteegh, and Emmanuel Dupoux. Learning words from images and speech. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2014.

R. Thiolliere, E. Dunbar, G. Synnaeve, M. Versteegh, and E. Dupoux. A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2015.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2017.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL `http://arxiv.org/abs/1807.03748`.

Balakrishnan Varadarajan, Sanjeev Khudanpur, and Emmanuel Dupoux. Unsupervised learning of acoustic sub-word units. In *Proceedings of ACL-08: HLT, Short Papers*, 2008.

Martin Versteegh, Roland Thiolliere, Thomas Schatz, Xuan Nga Cao, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux. The zero resource speech challenge 2015. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2015.

Virginia de Sa. Learning classification with unlabeled data. In *Proc. Neural Information Processing Systems (NeurIPS)*, 1994.

Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research (JMLR)*, 2009.

Yaodong Zhang and James Glass. Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2009.

Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2014.

A APPENDIX

A.1 VARYING THE CODEBOOK SIZE.

In Table 5, we examine the impact of varying the codebook size of model "$\varnothing \rightarrow \{2\}$" from 128 through 2048. We find that the ABX score is best for 1024 codebook vectors, although the performance is quite good for all models. Unsurprisingly, models with smaller codebooks also achieve lower bitrates.

Codebook size	R@10	ABX	Bitrate	RLE-Bitrate	Segment-ABX	Segment-Bitrate
128	.772	14.25	295.65	212.27	15.42	179.38
256	.756	12.95	341.18	260.10	14.21	228.07
512	.761	12.59	363.95	288.64	13.10	259.94
1024	.760	11.79	390.61	317.66	12.66	289.11
2048	.768	12.41	360.04	283.68	13.15	254.23

Table 5: ABX scores and bitrates for various codebook sizes on the clean ZeroSpeech19 English test set, using the "$\varnothing \rightarrow \{2\}$" model.

A.2 UNIT VISUALIZATION ON TIMIT SPECTROGRAMS

In Figure 5, we visualize two unit sequences for two different TIMIT utterances from different speakers containing the same underlying word sequence. Unit sequences are shown for the VQ2 layer for the "$\varnothing \rightarrow \{2\}$" model with a codebook size of 128.

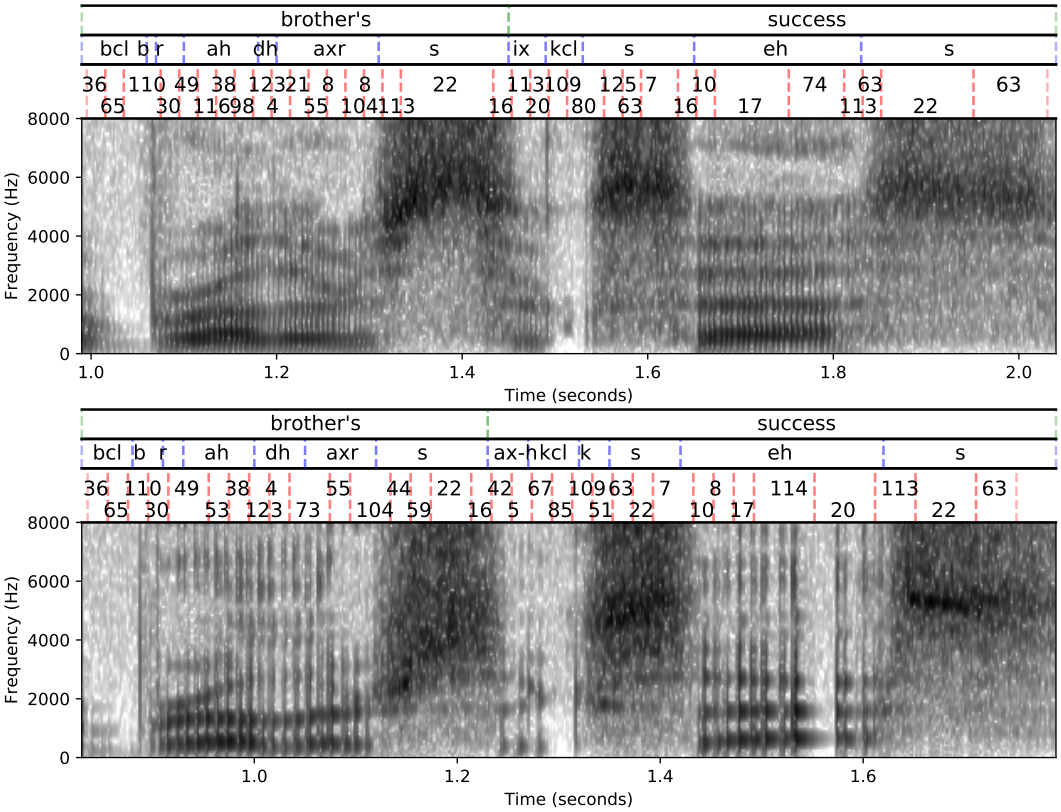

Figure 5: Two different instances of the same TIMIT sentence spoken by two different speakers. Notice how despite the differences between the speakers, the same units are often assigned to the same underlying phones.

A.3 VQ3 WORD DETECTOR TABLES FOR VARIOUS MODELS

In Table 6, we show a sampling of 50 word-detecting codebook entries from the VQ layer of the "$\{3\} \rightarrow \{2, 3\}$" model. Analagous results for the "$\{2\} \rightarrow \{2, 3\}$" model are shown in Table 7.

Table 6: Performance of the VQ3 layer from the "$\{3\} \rightarrow \{2, 3\}$" model when codes are treated as word detectors. Codes are ranked by the highest F1 score among the retrieved words for a given code. Word hypotheses for a given code are ranked by the F1 score.

rank	code	Top Hypotheses					Second Hypotheses				
		word	F1	P	R	occ	word	F1	P	R	occ
1	58	baseball	90.09	89.45	90.75	3266	player	2.82	1.60	12.18	139
2	706	background	89.26	92.36	86.36	17740	backgrounds	0.71	0.36	68.75	66
3	88	classroom	88.78	87.39	90.21	1512	class	4.65	2.79	13.99	82
4	116	construction	88.16	90.19	86.22	2484	constructed	1.98	1.03	26.29	56
5	584	playground	87.84	85.68	90.12	930	play	2.70	1.81	5.33	51
6	596	kitchen	87.26	86.42	88.11	5313	kitchenette	1.68	0.85	86.44	51
7	48	desert	87.17	87.85	86.50	3319	doesn't	1.59	0.89	7.41	28
8	625	background	86.79	94.12	80.52	16541	back	1.45	0.95	3.01	225
9	557	concrete	86.61	91.17	82.49	1917	country	0.64	0.49	0.92	13
10	5	airport	86.46	89.04	84.02	962	escalator	2.29	1.25	13.86	23
11	534	background	86.09	80.24	92.86	19076	back	8.44	5.67	16.52	1233
12	274	subway	86.01	86.62	85.41	1264	station	1.82	1.61	2.09	99
13	44	patio	85.93	89.94	82.27	2056	patios	2.02	1.02	77.78	28
14	310	rocky	85.90	86.19	85.62	2375	rock	3.52	2.72	4.97	266
15	18	driveway	85.68	91.60	80.49	1159	sidewalk	1.58	1.13	2.63	47
16	560	hospital	84.35	90.16	79.24	1191	hot	0.92	0.62	1.81	17
17	598	palm	83.68	82.26	85.16	2071	concrete	1.67	1.28	2.41	56
18	769	bamboo	83.68	85.57	81.88	1265	abandoned	5.38	3.31	14.44	109
19	892	walking	83.56	85.02	82.15	7747	walk	7.97	4.51	34.35	519
20	124	stage	83.43	84.65	82.24	2103	concert	1.60	0.87	10.74	51
21	802	body	83.39	84.24	82.56	6245	large	3.96	5.86	2.99	1324
22	162	stadium	83.08	80.35	86.00	1327	boardwalk	10.43	5.79	52.18	275
23	918	pantry	83.00	83.46	82.54	558	country	2.05	2.05	2.06	29
24	388	courtyard	82.88	88.76	77.74	1023	graveyard	2.91	1.66	11.72	45
25	641	volcano	82.81	79.72	86.15	336	volcanoes	5.11	2.66	65.00	13
26	85	yellow	82.78	78.72	87.28	11412	yellowish	2.38	1.21	84.49	158
27	1004	standing	82.78	85.87	79.90	12624	stand	5.08	2.84	24.38	534
28	630	trees	82.75	79.45	86.35	26838	tree	14.97	9.21	39.93	3436
29	148	course	82.74	86.86	79.00	1764	golf	6.54	6.89	6.23	195
30	661	distance	82.47	77.00	88.77	7779	background	1.93	1.82	2.06	423
31	394	church	82.42	74.32	92.51	4497	religious	5.97	3.19	47.73	263
32	108	highway	82.36	82.27	82.45	1038	highways	3.21	1.64	78.57	22
33	193	small	82.32	89.55	76.17	16417	smaller	2.18	1.12	39.55	263
34	326	shower	81.88	88.68	76.06	1312	showers	3.37	1.72	80.65	25
35	306	river	81.83	83.88	79.89	4393	rivers	1.09	0.55	60.98	25
186	280	station	68.23	54.23	91.95	4366	gas	28.72	17.35	83.40	1773
187	829	shirt	68.21	70.83	65.78	6592	shirts	15.44	8.70	68.57	757
188	548	night	68.15	64.95	71.69	2277	nighttime	13.14	7.15	81.42	241
189	554	computer	68.14	71.66	64.95	1360	computers	22.99	13.78	69.21	254
190	2	empty	67.96	76.93	60.87	2805	terminal	4.13	2.17	42.21	176
191	993	ruins	67.93	59.53	79.08	586	ruin	21.58	12.90	66.05	142
192	820	coffee	67.88	63.95	72.32	1335	cream	7.04	4.32	19.09	357
193	164	man	67.80	73.79	62.71	16622	men	13.85	9.77	23.79	1390
194	461	baby	67.71	64.00	71.88	1204	baby's	24.10	13.92	89.89	249
195	803	train	67.62	85.63	55.87	4720	trains	10.06	5.52	56.98	306
196	446	lake	67.56	78.98	59.02	2294	late	2.85	1.58	14.25	53
197	225	house	67.40	58.83	78.89	10776	houses	18.45	10.44	78.82	1868
198	234	alleyway	67.26	56.62	82.81	944	alley	37.25	30.06	48.98	863
199	1000	orange	67.25	89.09	54.00	2987	oranges	3.30	1.71	48.84	63
200	842	pink	67.19	78.12	58.94	2723	paint	5.05	3.07	14.26	140

Table 7: Performance of the VQ3 layer from the "$\{2\} \rightarrow \{2, 3\}$" model when codes are treated as word detectors. Codes are ranked by the highest F1 score among the retrieved words for a given code. Word hypotheses for a given code are ranked by the F1 score.

rank	code	Top Hypotheses					Second Hypotheses				
		word	F1	P	R	occ	word	F1	P	R	occ
1	924	people	76.60	66.79	89.80	33662	computer	2.30	1.18	42.36	887
2	530	building	75.96	65.57	90.27	35229	buildings	23.44	13.44	91.70	7260
3	749	white	75.86	66.14	88.93	45097	one	5.22	3.15	15.33	3315
4	505	blue	58.47	46.53	78.67	21678	pool	9.13	4.97	56.87	2674
5	581	snow	58.30	42.59	92.38	9746	small	16.17	11.72	26.06	5617
6	778	building	53.38	37.93	90.05	35144	buildings	14.46	7.85	91.28	7227
7	144	with	49.44	41.91	60.29	69909	looking	5.94	3.21	39.15	6194
8	299	small	48.88	33.61	89.58	19308	snow	30.58	18.38	90.95	9595
9	550	large	46.00	31.22	87.40	38648	car	8.35	4.41	79.56	4446
10	76	trees	45.43	30.22	91.45	28423	tree	15.18	8.31	87.91	7565
11	831	water	42.01	27.59	88.06	24901	wall	18.10	10.18	81.56	10428
12	80	red	39.67	26.38	79.94	19978	bed	9.46	5.08	68.27	3527
13	1015	large	39.36	25.85	82.37	36423	car	6.50	3.40	74.64	4171
14	719	woman	38.39	24.68	86.37	13913	women	7.01	3.70	65.24	2329
15	816	water	38.00	24.37	86.24	24388	river	5.32	2.76	74.00	4069
16	614	people	37.71	25.01	76.59	28709	table	22.26	12.94	79.67	13128
17	457	sky	36.42	23.76	77.99	10752	skies	10.67	5.68	86.90	2580
18	536	man	34.14	21.12	89.09	23614	standing	8.90	4.77	66.21	10460
19	985	trees	34.12	22.22	73.46	22832	tree	11.13	6.04	71.19	6126
20	0	black	33.50	20.83	85.51	23603	background	21.75	12.64	77.69	15960
21	870	trees	33.41	21.18	79.07	24577	train	13.53	7.34	87.36	7380
22	243	front	33.31	21.81	70.48	19397	from	14.71	8.72	47.02	7540
23	480	has	33.15	23.13	58.46	23919	house	9.95	5.56	47.15	6441
24	968	picture	32.94	21.29	72.74	33055	pictures	15.88	8.83	78.99	13439
25	245	yellow	32.91	20.16	89.55	11709	flowers	14.94	8.17	87.48	5623
26	815	with	32.65	22.54	59.20	68651	white	7.90	4.48	33.60	17037
27	153	yellow	32.25	19.76	87.63	11458	area	14.93	9.18	40.10	6956
28	526	small	32.21	19.65	89.32	19251	large	5.41	3.64	10.55	4664
29	293	large	31.95	22.88	52.92	23401	bridge	18.38	10.39	79.22	7066
30	538	black	31.32	19.53	78.98	21801	glass	8.57	4.52	81.80	4787
31	395	picture	30.79	20.31	63.58	28894	pictures	13.13	7.28	66.27	11276
32	971	man	29.89	18.58	76.36	20239	many	17.62	10.08	69.85	10692
33	133	white	29.80	18.86	71.02	36014	black	16.84	9.60	68.17	18818
34	740	trees	29.39	18.56	70.52	21918	station	3.72	1.91	66.22	3144
35	715	white	29.09	19.41	58.11	29465	like	9.73	5.75	31.49	7730
186	374	there's	6.42	3.62	28.70	14948	that	4.92	2.80	20.53	11745
187	869	old	6.41	3.39	58.22	6333	all	5.19	2.80	36.29	5490
188	982	structure	6.39	11.15	4.48	241	kids	2.62	1.98	3.88	55
189	746	has	6.25	3.44	33.94	13888	that	5.34	3.01	23.72	13565
190	791	with	6.15	3.78	16.43	19054	there's	3.15	1.76	14.96	7790
191	522	road	6.11	3.18	77.10	5614	field	5.93	3.11	63.57	7647
192	181	trees	6.11	3.28	45.19	14044	with	5.40	3.39	13.33	15456
193	790	photo	6.04	3.25	41.89	4958	by	5.51	2.99	35.62	4540
194	42	man	6.01	3.23	42.78	11340	middle	3.70	1.91	60.91	6059
195	801	sitting	5.82	3.10	48.26	8202	with	4.97	3.66	7.77	9009
196	975	purple	5.76	3.02	63.14	2388	parked	5.18	2.69	67.98	2567
197	432	parked	5.67	2.94	78.60	2968	park	5.11	2.64	82.26	2328
198	329	standing	5.29	2.80	49.17	7769	woman	3.15	1.65	35.41	5704
199	362	white	5.27	3.05	19.49	9883	table	4.71	2.49	44.61	7351
200	256	trees	5.24	2.88	28.84	8963	see	4.48	2.45	26.12	6346