Text-Free Image-to-Speech Synthesis Using Learned Segmental Units

Wei-Ning Hsu[1*], **David Harwath**[2*], **Tyler Miller**[2*], **Christopher Song**[3], **James Glass**[1]
[1]Massachusetts Institute of Technology
[2]University of Texas at Austin
[3]Johns Hopkins University
wnhsu@csail.mit.edu, harwath@utexas.edu

Abstract

In this paper we present the first model for directly synthesizing fluent, natural-sounding spoken audio captions for images that does not require natural language text as an intermediate representation or source of supervision. Instead, we connect the image captioning module and the speech synthesis module with a set of discrete, sub-word speech units that are discovered with a self-supervised visual grounding task. We conduct experiments on the Flickr8k spoken caption dataset in addition to a novel corpus of spoken audio captions collected for the popular MSCOCO dataset, demonstrating that our generated captions also capture diverse visual semantics of the images they describe. We investigate several different intermediate speech representations, and empirically find that the representation must satisfy several important properties to serve as drop-in replacements for text.

1 Introduction

Although there are over 7,000 languages spoken worldwide (Lewis et al., 2016), only several dozen have enough data available to support supervised speech recognition, and many languages do not even employ a writing system (Adda et al., 2016). In contrast, most people learn to use spoken language long before they learn to read and write, suggesting that linguistic annotation is not a prerequisite for speech processing systems. This line of reasoning motivates research that aims to discover meaningful linguistic abstractions (phones, words, etc.) directly from the speech signal, with the intention that they could reduce the reliance of spoken language systems on text transcripts.

A rich body of work has recently emerged investigating representation learning for speech using visual grounding objectives (Synnaeve et al., 2014; Harwath and Glass, 2015; Harwath et al., 2016; Kamper et al., 2017; Havard et al., 2019a; Merkx et al., 2019; Chrupała et al., 2017; Alishahi et al., 2017; Scharenborg et al., 2018; Hsu and Glass, 2018a; Kamper et al., 2018; Surís et al., 2019; Ilharco et al., 2019; Eloff et al., 2019), as well as how word-like and subword-like linguistic units can be made to emerge within these models (Harwath and Glass, 2017; Harwath et al., 2019; Drexler and Glass, 2017; Alishahi et al., 2017; Harwath et al., 2019; Harwath and Glass, 2019; Havard et al., 2019b; Harwath et al., 2020). So far, these efforts have predominantly focused on *inference*, where the goal is to learn a mapping from speech waveforms to a semantic embedding space. *Generation* of speech conditioned on a point in a semantic space has been less explored, and is what we focus on in this work. We hypothesize that generative approaches offer interesting advantages over relying solely on inference. For example, prior works have demonstrated the capability of recognizing visually descriptive words, but have not been shown to learn non-visual words or grammar. Our experiments show that these aspects of spoken language are learned to some degree by a visually-grounded generative model of speech.

Specifically, we introduce a model capable of directly generating fluent spoken audio captions of images without the need for natural language text, either as an intermediate representation or a form of supervision during training (Figure 1). Tremendous progress has been made recently in natural language image caption generation (Kiros et al., 2014; Mao et al., 2015; Vinyals et al., 2015; Karpathy and Fei-Fei, 2015; Xu et al., 2015; Rennie et al., 2017; Dai and Lin, 2017; Lu et al., 2017; Anderson et al., 2018; Lu et al., 2018) and naturalistic text-to-speech synthesis (TTS) (Ping et al., 2017; Taigman et al., 2017; Wang et al., 2017; Shen et al., 2018; Oord et al., 2016). Combining these models provides a means for generating spoken image descrip-

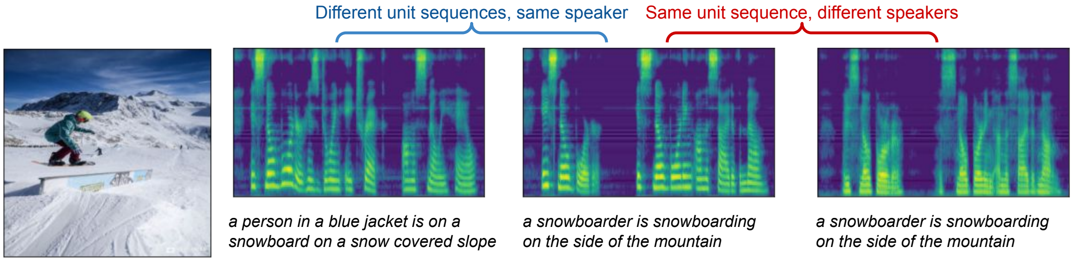

Different unit sequences, same speaker
Same unit sequence, different speakers
a person in a blue jacket is on a snowboard on a snow covered slope
a snowboarder is snowboarding on the side of the mountain
a snowboarder is snowboarding on the side of the mountain

Figure 1: Spoken image captions generated from the proposed model, with diversity in both linguistic content and acoustic properties, controlled through the I2U and the U2S models, respectively. Transcriptions are provided only for illustration. Audio samples are available at `https://wnhsu.github.io/image-to-speech-demo`.

tions, but existing approaches for training these models are reliant on text during training. Instead, we leverage sub-word speech units discovered using a self-supervised learning objective as a drop-in replacement for the text. We hypothesize that by using such techniques, an even wider variety of traditionally text-based NLP models could be applied to speech data without the need for transcription or automatic speech recognition (ASR) systems. Because all human languages utilize small, discrete phonetic inventories (International Phonetic Association, 1999), we posit that our framework should be applicable for any language in the world. In our experiments, we demonstrate that not just any set of discovered speech units can function in this role. We find the greatest success with units that are *discrete*, exhibit a *low frame-rate*, and *highly robust* to speaker and environmental variability. The main contributions of our paper are as follows:

1. The first methodology for fluent image-to-speech synthesis that does not rely on text. A critical aspect of our approach is factorizing the model into an Image-to-Unit (I2U) module and a Unit-to-Speech (U2S) module, where the speech units are discovered in a self-supervised fashion. This approach enables disentanglement of linguistic variability and acoustic/speaker variability.

2. Extensive analysis on the properties required for learned units to replace text. While the idea may seem simple and straightforward, obtaining proper units is not a trivial task. In fact, most of the units experimented in this paper fail to serve as drop-in replacements. Moreover, we demonstrate that what are deemed good units vary significantly for inference and generation.

3. Demonstrating insufficiency of beam search-based evaluation. We show that even when an I2U model fails to generate sensible caption through beam search decoding, it can still produce reasonable captions by sampling from the posterior, hinting that posterior mode-based evaluation can only inspect limited aspects of a model.

4. Proposing a semantic diversity-aware metric. We identify issues of an existing metric (Vijayakumar et al., 2018) and propose M-SPICE for sampling-based evaluation to address the problems.

5. Over 600,000 spoken audio captions for the MSCOCO dataset. We collect 742 hours of speech from 2,352 people tasked with reading each caption out loud. This dataset will be made publicly available to support work at the intersection of speech, language, and vision.

2 Related Work

Image-to-Text and Image-to-Speech Captioning. Significant progress towards generating realistic (text) captions that describe the content of visual images was made with the advent of deep neural networks (Vinyals et al., 2015; Karpathy and Fei-Fei, 2015; Xu et al., 2015; Anderson et al., 2018). Far less work has focused on generating spoken audio captions from natural images. Training an image-to-speech system using separate $(image, text)$ and $(text, speech)$ datasets was explored in (Ma et al., 2019). Hasegawa-Johnson et al. (2017) is the only prior work that has explored image-to-speech synthesis without using text, but with limited results. In that work, BLEU scores were only computed in terms of unsupervised acoustic units, not an estimate of the actual words produced by the synthesizer, which can be problematic as discussed in Section 4. The resulting captions were not evaluated for fluency, naturalness, or intelligibility, and the BLEU scores in terms of the unsupervised units were very low (0.014 on the MSCOCO test set) compared to ours (0.274). Wang et al. (2020b) is a concurrent work that proposes a text-free end-to-end image-to-speech model, which simplifies the task by using pairs of image and *synthesized* speech generated from a single-speaker TTS model to reduce the acoustic variation. In contrast, by leveraging robust learned units, our I2U module can be trained on real speech with abundant variation, and the U2S module serves as a vocoder that requires a small amount of clean speech (transcripts not needed). Hence, our system imposes less data constraints yet still outperforms Wang et al. (2020b).

Voice Conversion without Text aims to convert the speaker identity in a recording while preserving the textual content (Abe et al., 1990; Stylianou et al., 1998; Toda et al., 2007). It has recently seen progress using neural approaches (Hsu et al., 2016, 2017a,b; Fang et al., 2018; Chorowski et al., 2018; Chou et al., 2018; Lorenzo-Trueba et al., 2018; Serrà et al., 2019), but the most relevant work to our own is the ZeroSpeech 2019 challenge (Dunbar et al., 2019; Tjandra et al., 2019; Cho et al., 2019), which addresses unsupervised learning of discrete speech units that can replace text and be used as input to TTS models. Unlike image-to-speech synthesis, these tasks only *infer* phonetic units from given audio recordings instead of *generating* ones.

Speech Pre-Training and Its Applications. Interest in this area has recently surged. Various learning objectives have been proposed, including autoencoding with structured latent spaces (van den Oord et al., 2017; Eloff et al., 2019; Chorowski et al., 2019; Hsu et al., 2017b; Hsu and Glass, 2018b; Khurana et al., 2019), predictive coding (Chung et al., 2019; Wang et al., 2020a), contrastive learning (Oord et al., 2018; Schneider et al., 2019), and more. Prior work addresses inferring linguistic content such as phones from the learned representations (Baevski et al., 2020; Kharitonov et al., 2020; Hsu et al., 2021). In contrast, this work focuses on generating the learned representation from a different modality, which evaluates representations from a different perspective.

3 Method

3.1 Framework Overview

A depiction of our modeling approach is shown in Figure 2. Caption generation for an image involves a cascade of two components: given an input image I, we first generate a linguistic unit sequence U according to the I2U module $P(U \mid I)$. Given the linguistic symbol sequence U, we generate a speech waveform S according to the U2S module $P(S \mid U)$. If the linguistic unit sequence U were to take the form of natural language text, the model would be equivalent to the cascade of a conventional image captioning system followed by a TTS module. Note that we assume $S \perp I \mid U$ because prosody variation is not dependent on the image for the datasets considered.

The key idea in this paper is to instead define U to be a sequence of *learned* speech units that are as *robust and compact* as possible like text, but discovered without text supervision. We define inference with this S2U model as $U = f(S)$, enabling us to "transcribe" any given speech audio waveform S into a sequence of units U. The addition of this third component enables us to train $P(U \mid I)$ from a dataset of images paired with spoken captions $\{(I_1, S_1), \ldots, (I_N, S_N)\}$. The conditional independence assumption between S and I given the U enables us to choose any arbitrary speech dataset for training $P(S \mid U)$, therefore enabling the speaker characteristics and other acoustic properties to be independently controllable from the I2U system (Wang et al., 2018; Hsu et al., 2019; Henter et al., 2018; Akuzawa et al., 2018).

3.2 Datasets

Table 1 summarizes the five datasets used for training S2U, I2U, and U2S models. Note that we deliberately choose different datasets for training each module, which aims to examine the robustness of the units when transferring across domains, including shift in speaker demography, speaking style (scripted/spontaneous), and linguistic content (book/newspaper/image description). Among the three datasets with image and speech pairs: Places, Flickr8k, MSCOCO, we chose the latter two for training I2U models, because they include five captions per image, which is more suitable for caption metrics such as SPICE (Anderson et al., 2016); moreover, they are commonly used image captioning datasets with many text-based baselines in the literature. Places only contains one spoken caption per image and has not been used for captioning.

Specifically, as part of this work we collect **SpokenCOCO**, a spoken version of the MSCOCO captioning dataset (Lin et al., 2014) with 742 hours from 2532 speakers, via Amazon Mechanical Turk by displaying the text to a person and having them read it aloud. Additional details regarding the dataset can be found in appendix Section A. Note that although there exists a speech version

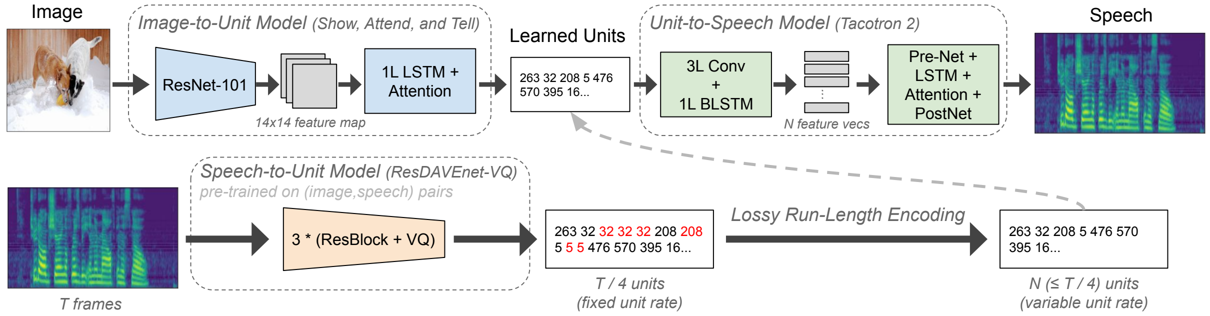


Figure 2: Diagram of our proposed framework. The ResDAVEnet-VQ model was trained using a $\{2\} \rightarrow \{2,3\}$ curriculum (in the notation given in Harwath et al. (2020)).

	Data	Hr	#Utt	#Spk	Maj. Spk	Description
S2U	PlacesAudio (Harwath et al., 2016)	936	400K	2683	American	spontaneous image caption
I2U	Flickr8kAudio (Harwath and Glass, 2015) SpokenCOCO (this work)	46 742	40K 605K	183 2353	American	scripted image caption
U2S	LJSpeech (Ito, 2017) VCTK (Veaux et al., 2017)	24 44	13K 40K	1 109	American British	read non-fiction books read newspaper

Table 1: Speech dataset summary. For training S2U and I2U models, their corresponding image datasets, MSCOCO (Lin et al., 2014), Flickr8k (Rashtchian et al., 2010), and Places (Zhou et al., 2014), are also used.

of MSCOCO named Speech-COCO (Havard et al., 2017), it is comprised of only synthesized speech using a concatenative TTS model in eight speakers' voice. Disfluencies (e.g. "uh") are randomly inserted in between words to imitate real speech. Compared to SpokenCOCO, Speech-COCO offers limited diversity and naturalness.

3.3 Learning Robust Linguistic Units from Visually-Grounded Speech

We propose to build the S2U model upon ResDAVEnet-VQ, an audio-visual grounding model introduced in Harwath et al. (2020) that has shown to learn discrete phone- and word-like units in the intermediate vector quantizing (VQ) layers. This model is trained to associate speech with contextually relevant visual inputs using a triplet loss (Weinberger and Saul, 2009), which can be interpreted as maximizing a mutual information lower bound between image and speech (Tschannen et al., 2020). Since visual semantics are described with words, which in turn are composed of phones, the representations learned by ResDAVEnet-VQ are forced to be predictive of words and phones rather than speaker, noise, etc.

In contrast, many of the speech representations are trained by reconstructing (Chorowski et al., 2019; Hsu et al., 2017b) or predicting unseen speech signals (Chung et al., 2019), which would inevitable capture factors unrelated to the linguistic content. To demonstrate the advantage of representation learning with grounding, we will compare ResDAVEnet-VQ with a reconstruction based model, WaveNet-VQ, trained on the PlacesAudio dataset. We denote the units extracted from this model with WVQ. We use the implementation of Harwath et al. (2020) for ResDAVEnet-VQ, and Cho et al. (2019) for WaveNet-VQ which achieves the best ZeroSpeech 2019 challenge performance.

3.4 Unit Selection and Run Length Encoding

Although the ResDAVEnet-VQ model has been shown to be capable of learning both phone-like and word-like units, the experiments in (Harwath et al., 2020) show that only several hundred words are explicitly learned, which tend to be "visual words." Conversely, the phone-like units learned by the lower VQ layers of the model were shown to cover all of the phones in American English (as there are only several dozens). For this reason, we choose to use phone-like units learned by the lower VQ layers to represent U.

Nominally, the VQ layers will output one-hot vectors at a uniform temporal rate, downsampled with respect to the framerate of the acoustic input depending upon which VQ layer is used. Given an input computed with a 10ms frame shift, the two VQ layers investigated in this paper (VQ2 and VQ3) respectively output vectors every 20ms and 40ms. In general, the VQ units are repeated for several consecutive frames. We can decrease the average length of the symbol sequence U by employing a lossy form of **run-length encoding** (RLE) (see Figure 2) which retains the sequence of symbol identities but discards duration information. Each unit then represents a variable-length segment. This removes the burden of unit duration modeling from the I2U model and shifts it onto the U2S model, which we will show to be crucial.

3.5 Image-to-Unit and Unit-to-Speech

Both the I2U model and the U2S model are based upon recurrent seq2seq with attention networks (Bahdanau et al., 2015). Specifically, we adopt Show-Attend-and-Tell (SAT) (Xu et al., 2015) for the I2U model. It has an image encoder pre-trained for classification, which is language agnostic and hence should work in any language within our proposed framework. The decoder on the other hand is randomly initialized. We train the SAT model for two stages, where the encoder parameters are only updated in the second stage. We distinguish the models from the two stages with *SAT* and *SAT-FT* (finetuned) respectively when presenting the results. For the U2S model, we adopt Tacotron2 (Shen et al., 2018) and WaveGlow (Prenger et al., 2019) for unit-to-spectrogram and spectrogram-to-waveform generation, respectively. In particular, a pre-trained WaveGlow is used without fine-tuning.

The I2U model is trained on $(I, f(S))$ pairs, which requires pairs of image and speech, while the U2S model is trained on $(f(S), S)$ pairs, which can be obtained from arbitrary set of speech. Both models are trained with the maximum likelihood objective ($\mathbb{E}_{I,U}\left[\log P(U \mid I)\right]$ for I2U and $\mathbb{E}_{S,U}\left[\log P(S \mid U)\right]$ for U2S).

4 Experiments

We design experiments to address three questions:

First, how can we measure the performance of an image-to-speech system? Our system can fail to produce a good caption if the I2U model fails to encode linguistic/semantic information into the unit sequence, or if the U2S model fails to synthesize an intelligible waveform given a unit sequence. To better localize these failure modes, we evaluate the full I2S system as well as the U2S system in isolation. We evaluate the U2S system by using it as a vocoder to synthesize unit sequences *inferred from real speech* and soliciting human judgements in the form of Mean Opinion Score (MOS) and Side-By-Side (SXS) preference tests (Table 2).

To evaluate the I2S system, we can use any method that measures the semantic information contained in the generated speech. We consider two sets of end-to-end metrics: *word-based* and *retrieval-based*, and one set of proxy *unit-based* metrics. Word-based metrics transcribe a generated spoken caption into text (manually or with an ASR system) and then measure word-based captioning metrics against a set of reference captions, such as BLEU-4 (Papineni et al., 2002) (adjusted n-gram precision), METEOR (Denkowski and Lavie, 2014) (uni-gram F-score considering word-to-word alignment), ROUGE (Lin, 2004) (n-gram recall), CIDEr (Vedantam et al., 2015) (TF-IDF weighted n-gram cosine similarity), and SPICE (Anderson et al., 2016) (F-score of semantic propositions in scene graphs). This enables comparison between image-to-speech systems with a text "upperbound", but is not applicable to unwritten languages.

Retrieval-based metrics include image-to-speech and speech-to-image retrieval (Harwath et al., 2020), which require a separately trained cross-modal retrieval model for evaluation. Such metrics are text-free, but they cannot measure other aspects of language generation such as syntactic correctness (partially captured by BLEU-4) and scope of the learned vocabulary. Lastly, unit-based metrics are similar to text-based, but replace words with units when computing n-gram statistics. However, systems built on different units are not directly comparable, and second, can be inflated if duration is modeled using unit repetition.

Second, what properties must learned units have to be a drop-in replacement for text? The most essential differences between text and speech are the amount of information encoded and the sequence lengths. Beyond text, speech also encodes prosody, speaker, environment information and the duration for each phone, all of which are minimally correlated with the conditioned images. We hypothesize that learned speech units should discard such information in order to seamlessly connect the I2U and U2S modules. To verify it, we pay particular attention to the variations of the learned units in *frame rate* (VQ2/VQ3), *encoding of duration information* (RLE or not), and *robustness to domain shift* (WVQ/VQ3). Units are run-length encoded by default. Table 2a shows the properties of the units before run-length encoding.

Third, how should language generation models be evaluated more generally? We examine evaluation of the I2S model using beam search-based decoding as well as sampling-based decoding. We find that because evaluation metrics that are reliant on beam search-based decoding only evaluate the *mode* of a model's posterior, they do not reflect the ability of a model to generate diverse linguistic content. Furthermore, we show that it is possible for a model's posterior mode to be linguistically meaningless, and yet meaningful language can still be generated with sampling-based decoding. Towards this end, we introduce a novel multi-hypothesis evaluation metric (M-SPICE), which uses sampling-based decoding (instead of beam search) to generate a set of captions. We can then compute the overall coverage of this caption set against a reference; see Section 4.4 for details.

4.1 Evaluating the U2S Model

We construct a Tacotron-2 model for each of the three unit types on the LJSpeech audio data by transcribing each LJSpeech utterances into an unit sequence, then train the U2S model from the RLE-ed unit sequence and spectrogram pairs. We evaluate the naturalness of the speech produced by each model on held-out data, both in-domain using LJSpeech and out-of-domain (OOD) using SpokenCOCO.[1] Amazon Mechanical Turk (AMT) workers performed Side-by-Side preference tests (SXS) and naturalness evaluation based on mean opinion scores (MOS) on a scale from 1 to 5 for each U2S model, which we display in Table 2. Although VQ2 was preferred for in-domain synthesis on LJSpeech, VQ3 achieved the highest scores and least degradation (-0.387) on the out-of-domain SpokenCOCO, indicating that out of the three units VQ3 has the strongest robustness to domain shift.

Unit	ABX Error	Frame Rate	MOS LJSpeech	MOS SpokenCOCO
VQ3	14.52%	40ms	3.723 ± 0.039	3.336 ± 0.044
VQ2	12.51%	20ms	3.932 ± 0.036	2.961 ± 0.045
WVQ	24.87%	40ms	3.658 ± 0.040	2.896 ± 0.053

(a) Properties of the units and MOS of the U2S models trained on these units with 95% confidence interval. ABX errors are computed on the ZeroSpeech 2020 English test set.

Unit		LJSpeech			SpokenCOCO		
A	B	A	Same	B	A	Same	B
VQ3	VQ2	23.9	31.5	44.6	40.4	32.5	27.5
VQ3	WVQ	36.6	37.1	26.3	58.3	21.8	19.9

(b) SXS preference (%) of the U2S models.

Table 2: Subjective evaluation of U2S models trained on LJSpeech and re-synthesize units inferred from LJSpeech or SpokenCOCO recordings.

4.2 Incorporating the I2U Model

We trained an SAT model on SpokenCOCO for each of the three RLE-ed units, as well as VQ3 units without RLE. We also compare to text characters and words; the full hyperparameter and training details for all models are provided in Section B in the appendix, but in general we kept these as constant as possible when comparing different linguistic representations.

Symbol	Image-to-Unit Output
Decoded with Beam Search (beam size=5)	
VQ3	263 32 208 5 336 100 717 803 256 803 815 144 120 144 654 936 48 417 272 417 362 766 825 284 614...
VQ2	(71 791)*N (until reaching max decoder length)
WVQ	(181 232)*N (until reaching max decoder length)
VQ3 \ RLE	263 (32)*N (until reaching max decoder length)
Decoded with Top-k Sampling Search (*k*=5)	
VQ3	263 208 467 717 288 426 986 72 44 341 151 801 1022 27 320 426 288 66 570 683 351 313 910 820...
VQ2	(71 791)*4 175 51 139 359 173 599 307 419 133 621 85 165 315 883 175 191 71 791 71 48 511 765...
WVQ	(181 232)*5 181 225 124 232 181 232 225 232 181 225 124 225 232 181 252 169 211 147 89 67 156...
VQ3 \ RLE	263 (32)*15 208 208 5 5 336 100 803 256 560 417 870 870 870 968 910 250 543 820 587 909 909...

Table 3: Exemplar output from SAT models.

Before connecting the U2S model, we noticed that all RLE speech unit models except the one trained on VQ3 units failed during *beam search* decoding on the test images (WVQ consistently failed, while VQ2 sometimes succeeded); rather than producing a diverse sequence of output units, the decoder would generally get stuck in a loop until the maximum decoding length was reached. This also happened using VQ3 units without RLE, indicating that the decoder could not model unit duration. Example outputs are provided in Table 3. We hypothesize that the reason the VQ2 and WVQ units failed is due to their lack of invariance to domain shift, as evidenced by their decay in naturalness when used for OOD synthesis as shown in Table 2. This may cause the entropy of the unit distribution conditioned on an image to be higher as each phoneme may be represented by multiple units, and therefore the I2U model suffers from the same looping issues as the unconditional language model of text, as observed in (Holtzman et al., 2018; Fan et al., 2018; Holtzman et al., 2020;

[1] In-domainess is defined with respect to the U2S training data (LJSpeech) not the S2U training data (PlacesAudio).

model	U	MSCOCO					Flickr8k				
		B-4	M	R	C	S	B-4	M	R	C	S
Xu et al. (2015)	word	0.243	0.239	-	-	-	0.213	0.203	-	-	-
Lu et al. (2017)	word	0.327	0.260	0.540	1.042	-	-	-	-	-	-
Wang et al. (2020b)	N/A	-	-	-	-	-	0.035	0.113	0.232	0.080	-
SAT	word	0.315	0.253	0.533	0.984	0.185	0.216	0.207	0.469	0.550	0.149
	char	0.289	0.239	0.512	0.879	0.172	0.190	0.190	0.441	0.476	0.136
	VQ3	0.186	0.186	0.446	0.584	0.127	0.116	0.141	0.390	0.232	0.091
SAT-FT	word	0.339	0.265	0.551	1.062	0.196	0.225	0.215	0.483	0.584	0.155
	char	0.323	0.256	0.536	1.002	0.187	0.191	0.196	0.450	0.519	0.143
	VQ3	0.233	0.212	0.478	0.732	0.149	0.125	0.145	0.391	0.245	0.095

Table 4: Word-based caption evaluation using **B**LEU-**4**, **M**ETEOR, **R**OUGE, **C**IDEr, and **S**PICE. ASR is used to transcribe the spoken captions generated by the proposed VQ3 model into text for evaluation. The beam size $\in \{1, 3, 5, 8, 10\}$ was chosen for each model to maximize SPICE. Our word-based SAT models outperform (Xu et al., 2015) because we use a stronger image encoder (ResNet-101).

symbol	Word-based					Unit-based				Retrieval-based					
										Image to Speech			Speech to Image		
	B-4	M	R	C	S	B-4	M	R	C	R@1	R@5	R@10	R@1	R@5	R@10
SAT-FT Model, Decoded with Beam Search															
VQ3	0.233	0.212	0.478	0.732	0.149	0.261	0.198	0.334	0.211	0.240	0.603	0.766	0.265	0.611	0.765
SAT Model, Decoded with Beam Search															
VQ3	0.186	0.186	0.446	0.584	0.127	0.274	0.196	0.328	0.215	0.157	0.451	0.623	0.158	0.450	0.611
VQ2	0.068	0.138	0.343	0.262	0.084	0.172	0.132	0.178	0.027	0.09	0.289	0.426	0.093	0.283	0.420
WVQ	0.010	0.069	0.286	0.009	0.011	0.020	0.048	0.081	0.000	0.000	0.005	0.010	0.001	0.006	0.011
VQ3 \ RLE	0.000	0.002	0.001	0.000	0.001	0.163	0.168	0.218	0.000	0.000	0.003	0.007	0.001	0.006	0.011

Table 5: Comparison of the three sets of metrics on different units and models trained on MSCOCO.

Kulikov et al., 2019; Welleck et al., 2020).

To evaluate the full Image-to-Speech model, we first train an ASR system on the re-synthesized SpokenCOCO captions using the VQ3 Tacotron-2 model. This enables us to estimate a word-level transcription of the spoken captions produced by our system. In order to verify that the synthesized captions are intelligible to humans and the ASR system did not simply learn to recognize artifacts of the synthesized speech, we asked AMT workers to transcribe into words a set of 500 captions generated by our I2U→U2S system and also evaluated their naturalness. Three workers transcribed and three workers rated each caption, allowing us to compute an **MOS score (3.615±0.038)**, a word error rate (**WER**) between the 3 human transcriptions (9.40%), as well as an average WER between the human and ASR-produced transcriptions (**13.97%**). This confirms that our system produces reasonably natural speech and ASR is sufficiently accurate for transcribing synthesized speech.

Table 4 summarizes our results on MSCOCO and Flickr8k using beam search. We compare with the literature for bottom-up text captioning (row 1-2) and text-free end-to-end image-to-speech synthesis (row 3). We train the decoder of an SAT model while keeping the image encoder fixed (row 4-6), in addition to fine-tuning the encoder (row 7-9). Despite having no access to text, *the SAT-FT speech captioning model trained on VQ3 units achieves a BLEU-4 score of .233 with beam search decoding on MSCOCO. This is very close to the .243 achieved by the original SAT word-based captioning model.* Figure 1 shows that the generated captions are fluent and reflect the implicit learning of some syntactic rules. It is evident that the proposed model is capable of generating fluent and meaningful image captions.

Results comparing four unit representations on all three sets of metrics are shown in Table 5. First of all, by comparing word-based and unit-based evaluations, we do note that the relative ranking among VQ3, VQ2, and WVQ is consistent across BLEU-4, METEOR, and ROUGE for SAT models, however, VQ3 \ RLE achieves abnormally high scores on these metrics despite producing trivial captions for all images as shown in Table 3. This is because unit "32" has learned to represent non-speech frames such as silence, which frequently occurs at both the beginning and end of utterances. Without RLE, consecutive strings of "32" units are extremely common in both the candidate and reference captions, which inflates the scores of this model. The exception here is the CIDEr metric,

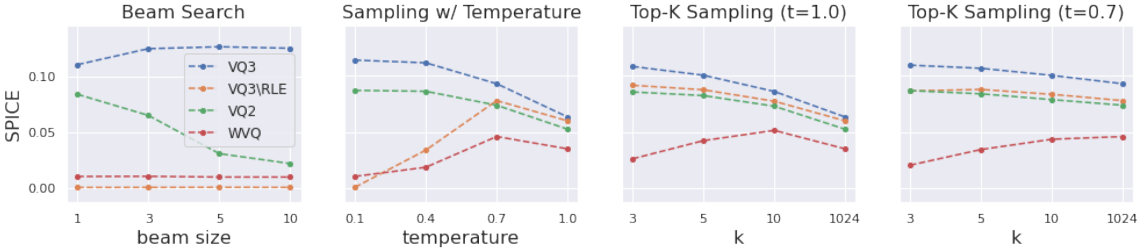


Figure 3: MSCOCO test SPICE scores of various units and decoding methods. VQ3\RLE denotes VQ3 units without RLE. Top-k sampling considers only the k-most probable units at each step.

which incorporates TF-IDF weighting that tends to de-emphasize these kinds of uninformative patterns. Nonetheless, when comparing SAT and SAT-FT with VQ3 units, CIDEr does not rank them the same as word-based metrics.

Regarding retrieval-based evaluation, despite the fact that the ResDAVEnet model was only trained on the original, human-spoken captions for the MSCOCO images, it works very well for the fully synthetic captions. The speech and image retrieval scores for 1k human-spoken validation captions are 0.867 and 0.828 R@10, respectively, while the SAT-FT VQ3 model achieves 0.766 and 0.765 R@10. This indicates that this image-to-speech model is able to infer the salient semantic content of an input image, generate a unit sequence that captures that content, and generate speech that is sufficiently natural sounding for the ResDAVEnet model to recover that semantic information. Several of the other image-to-speech models also achieve respectable retrieval performance, and the overall ranking of the models mirrors that which we found when using word-based evaluation metrics.

4.3 From Mode to Distribution: Evaluating Captions Generated via Sampling

The results in the previous section only evaluate beam search decoding with the I2U model, and do not fully reveal the posterior over captions for an input image, or whether the unit representations that failed with beam search would work well with other methods. To probe this, we evaluate the models using sampling-based caption generation. Figure 3 shows the SPICE scores on SpokenCOCO using beam search and two sampling-based methods. VQ3 still performs the best of all unit types with both beam search and sampled decoding. VQ2 can sometimes generate captions with beam search when the beam is kept small, but as the beam grows it begins to loop and the scores become very low.

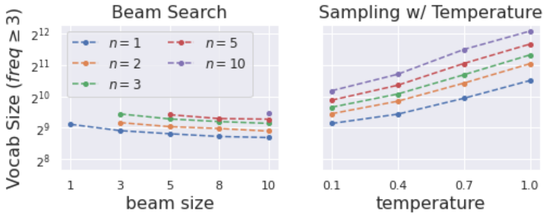


Figure 4: Vocabulary size learned by the proposed I2S model (on MSCOCO)

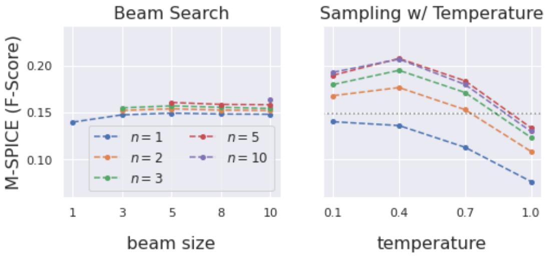


Figure 5: M-SPICE on MSCOCO. Black dashed lines show the highest value for beam search when n=1.

We see that all unit types can generate reasonable captions when decoding via sampling. Moreover, we discovered that 1) ResDAVEnet-VQ units consistently outperform the WaveNet-VQ units, suggesting that they better capture sub-word structure, and 2) VQ3 \ RLE achieves better scores than VQ2 when using a larger temperature or k for top-k.

We estimated the vocabulary size of the SAT-FT model with VQ3 by counting the number of unique recognized words produced at least 3 times when captioning the SpokenCOCO test images. These numbers are shown for the model under the various decoding methods in Figure 4. The number of captions per image is denoted by n, where top candidates are used for beam search and i.i.d. samples are drawn for sampling. Sampling-based decoding reveals a larger vocabulary size than beam search, and the number of words learned by our models ($\geq 2^{12}$) is far greater than the number of words learned by the ResDAVEnet-VQ model (approx.

Speaker	Gender	Region	B4	M	S
U2S trained on LJSpeech					
-	F	-	0.233	0.212	0.149
U2S trained on VCTK					
p247	M	Scottish	0.234	0.211	0.148
p231	F	English	0.233	0.210	0.146
p294	F	American	0.236	0.212	0.148
p345	M	American	0.234	0.209	0.144
p307	F	Canadian	0.234	0.211	0.148

Table 6: Results of disentangled voice control via synthesizing the same units with a single and a multi speaker U2S model. Units are decoded using beam search from the SAT-FT VQ3 MSCOCO model.

279) in (Harwath et al., 2020). We hypothesize that training a model to *generate* spoken captions encourages it to learn many more words than only being trained to retrieve images from captions. We also hypothesize that because beam search attempts to find the mode of the posterior over captions, it tends to produce a smaller set of words and does not reveal the breadth of the model distribution.

4.4 New Diversity-Aware Metric: M-SPICE

The previous section showed that even when the SPICE scores were comparable, sampling-based decoding revealed a much larger model vocabulary than beam search, especially when multiple captions are generated for each image. This highlights a limitation of SPICE in measuring the *diversity*. Formally speaking, SPICE computes an F-score between two bags of semantic propositions $T(S)$ and $T(c)$ parsed from a set of references $S = \{s_i\}_i$ and a hypothesis c, where $T(c)$ denotes a bag of propositions extracted from a scene graph parsed c, and we can compute that for multiple sentences with $T(S) = \cup_i(T(s_i))$.

To extend SPICE for scoring multiple hypotheses $C = \{c_j\}_{j=1}^J$, one can compute an average SPICE: $\frac{1}{J}\sum_j F1(T(S), T(c_j))$, or use the oracle SPICE proposed in Vijayakumar et al. (2018): $max_j F1(T(S), T(c_j))$. However, these metrics fail to capture the diversity *among* hypotheses. Consider two hypothesis set, $C^1 = \{c_1^1, c_2^1\}$ and $C^2 = \{c_1^2, c_2^2\}$, where $T(c_1^1) = T(c_2^1) = T(c_1^2) = \{$(girl), (table), (girl, sit-at, table)$\}$, $T(c_2^2) = \{$(girl), (girl, young)$\}$, and $T(S) = \{$(girl), (table), (girl, young), (girl, sit-at, table)$\}$.

To address the deficiencies of the existing metrics, we propose a new metric named multi-candidate SPICE (M-SPICE), which takes the *union of the candidate propositions* and computes the F-score against the reference propositions: $F1(T(S), \cup_j T(c_j))$. M-SPICE assigns a higher score if the set captures *diverse and correct* propositions, and it is obvious that the score of C^2 is higher than C^1 as desired.Figure 5 shows the M-SPICE scores of our SAT-FT model using VQ3 units on SpokenCOCO. When evaluating over multiple captions ($n > 1$), using the beam search hypotheses increases the score less than sampling.

4.5 Disentangled Voice Control for Image-to-Speech Synthesis

We examine to what extent the VQ3 units are portable across different speakers by training a U2S model on the VCTK dataset that additionally takes a speaker ID as input. The resulting model is able to generate speech with the voice of any VCTK speaker. We evaluate the captions produced by this system on SpokenCOCO for 5 speakers in Table 6. To compute these scores we transcribe the captions generated by each model into text using the ASR system we describe in Section 4.2, which was solely trained on re-synthesized SpokenCOCO captions using the LJSpeech U2S model. The scores in Table 6 indicate not only that the I2U model can be easily integrated with U2S models representing a diverse set of speakers, but also that the LJSpeech ASR system works very well on the speech synthesized from the VCTK models.

5 Conclusion

In this paper, we presented the first model capable of generating fluent spoken captions of images without relying on text, which almost matches the performance of early text-based image captioning models. Our comprehensive experiments demonstrated that learned units need to be robust, of low framerate, and encoding little or none duration information to be a drop-in replacement for text. We also identified the caveats of mode-based evaluation and proposed a new metric to address semantic diversity. As part of this work, a novel dataset of over 600k spoken captions for the MSCOCO dataset is introduced, which we will make publicly available to the research community.

Future work should investigate applying the proposed method to additional languages, devising improved speech unit representations, and jointly training the speech unit model with the I2S model. This would offer the opportunity to explore new analysis-by-synthesis training objectives.

References

Masanobu Abe, Satoshi Nakamura, Kiyohiro Shikano, and Hisao Kuwabara. 1990. Voice conversion through vector quantization. *Journal of the Acoustical Society of Japan*, 11(2):71–76.

Gilles Adda, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélene Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, et al. 2016. Breaking the unwritten language barrier: The BULB project. *Procedia Computer Science*, 81:8–14.

Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. 2018. Expressive speech synthesis via modeling expressions with variational autoencoder. *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*.

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

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic propositional image caption evaluation. In *Proc. IEEE European Conference on Computer Vision (ECCV)*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alexei Baevski, Steffen Schneider, and Michael Auli. 2020. vq-wav2vec: Self-supervised learning of discrete speech representations. In *Proc. International Conference on Learning Representations (ICLR)*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proc. International Conference on Learning Representations (ICLR)*.

Suhee Cho, Yeonjung Hong, Yookyunk Shin, and Youngsun Cho. 2019. VQVAE with speaker adversarial training.

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

Jan Chorowski, Ron J Weiss, Rif A Saurous, and Samy Bengio. 2018. On using backpropagation for speech texture generation and voice conversion. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Ju-chieh Chou, Cheng-chieh Yeh, Hung-yi Lee, and Lin-shan Lee. 2018. Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*.

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

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

Bo Dai and Dahua Lin. 2017. Contrastive learning for image captioning. In *Proc. Neural Information Processing Systems (NeurIPS)*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *In Proceedings of the Ninth Workshop on Statistical Machine Translation*.

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

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. 2019. The zero resource speech challenge 2019: TTS without T. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*.

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

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*.

Fuming Fang, Junichi Yamagishi, Isao Echizen, and Jaime Lorenzo-Trueba. 2018. High-quality nonparallel voice conversion based on cycle-consistent adversarial network. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

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

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

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

David Harwath, Wei-Ning Hsu, and James Glass. 2020. Learning hierarchical discrete linguistic units from visually-grounded speech. In *Proc. International Conference on Learning Representations (ICLR)*.

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

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

Mark Hasegawa-Johnson, Alan Black, Lucas Ondel, Odette Scharenborg, and Francesco Ciannella. 2017. Image2speech: Automatically generating audio descriptions of images. In *International Conference on Natural Language, Signal and Speech Processing*.

William Havard, Laurent Besacier, and Olivier Rosec. 2017. Speech-coco: 600k visually grounded spoken captions aligned to mscoco data set. *arXiv preprint arXiv:1707.08435*.

William Havard, Jean-Pierre Chevrot, and Laurent Besacier. 2019a. 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)*.

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

Gustav Eje Henter, Jaime Lorenzo-Trueba, Xin Wang, and Junichi Yamagishi. 2018. Deep encoder-decoder models for unsupervised learning of controllable speech synthesis. *arXiv preprint arXiv:1807.11470*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proc. International Conference on Learning Representations (ICLR)*.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*.

Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang. 2016. Voice conversion from non-parallel corpora using variational auto-encoder. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*.

Wei-Ning Hsu and James Glass. 2018a. Disentangling by partitioning: A representation learning framework for multimodal sensory data. *arXiv preprint arXiv:1805.11264*.

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

Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: How much can a bad teacher benefit asr pre-training? In *ICASSP*. IEEE.

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

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

Wei-Ning Hsu, Yu Zhang, Ron Weiss, Heiga Zen, Yonghui Wu, Yuan Cao, and Yuxuan Wang. 2019. Hierarchical generative modeling for controllable speech synthesis. In *Proc. International Conference on Learning Representations (ICLR)*.

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

International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

Keith Ito. 2017. The LJ speech dataset. `https://keithito.com/LJ-Speech-Dataset/`.

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

Herman Kamper, Gregory Shakhnarovich, and Karen Livescu. 2018. Semantic speech retrieval with a visually grounded model of untranscribed speech. *IEEE Transactions on Audio, Speech and Language Processing*, PP:1–1.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Eugene Kharitonov, Morgane Rivière, Gabriel Synnaeve, Lior Wolf, Pierre-Emmanuel Mazaré, Matthijs Douze, and Emmanuel Dupoux. 2020. Data augmenting contrastive learning of speech representations in the time domain. *arXiv preprint arXiv:2007.00991*.

Sameer Khurana, Shafiq Rayhan Joty, Ahmed Ali, and James Glass. 2019. A factorial deep markov model for unsupervised disentangled representation learning from speech. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations (ICLR)*.

Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In *Proc. International Conference on Machine Learning (ICML)*.

Ilia Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. 2019. Importance of search and evaluation strategies in neural dialogue modeling. In *Proceedings of the 12th International Conference on Natural Language Generation*.

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

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. *ArXiv*, abs/1405.0312.

Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling. 2018. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. *arXiv preprint arXiv:1804.04262*.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shuang Ma, Daniel McDuff, and Yale Song. 2019. Unpaired image-to-speech synthesis with multimodal information bottleneck. In *Proc. IEEE International Conference on Computer Vision (ICCV)*.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). In *Proc. International Conference on Learning Representations (ICLR)*.

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

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

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2017. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *arXiv preprint arXiv:1710.07654*.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. 2019. Waveglow: A flow-based generative network for speech synthesis. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proc. NAACL Conference on Human Language Technologies (NAACL-HLT)*.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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. 2018. 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)*.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*.

Joan Serrà, Santiago Pascual, and Carlos Segura Perales. 2019. Blow: a single-scale hyperconditioned flow for non-parallel raw-audio voice conversion. In *Proc. Neural Information Processing Systems (NeurIPS)*.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE.

Yannis Stylianou, Olivier Cappé, and Eric Moulines. 1998. Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*, 6(2):131–142.

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

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

Yaniv Taigman, Lior Wolf, Adam Polyak, and Eliya Nachmani. 2017. Voiceloop: Voice fitting and synthesis via a phonological loop. *arXiv preprint arXiv:1707.06588*.

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 preprint arXiv:1905.11449*.

Tomoki Toda, Alan W Black, and Keiichi Tokuda. 2007. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech and Language Processing*, 15(8):2222–2235.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. 2020. On mutual information maximization for representation learning. In *Proc. International Conference on Learning Representations (ICLR)*.

Christophe Veaux, Junichi Yamagishi, and Kirsten Macdonald. 2017. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasaath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Weiran Wang, Qingming Tang, and Karen Livescu. 2020a. Unsupervised pre-training of bidirectional speech encoders via masked reconstruction. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Xinsheng Wang, Siyuan Feng, Jihua Zhu, Mark Hasegawa-Johnson, and Odette Scharenborg. 2020b. Show and speak: Directly synthesize spoken description of images. *arXiv preprint arXiv:2010.12267*.

Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif Saurous. 2017. Tacotron: Towards end-to-end speech synthesis. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *arXiv preprint arXiv:1803.09017*.

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

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *Proc. International Conference on Learning Representations (ICLR)*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. International Conference on Machine Learning (ICML)*.

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

A Visually-Grounded Speech Datasets

Table A1 displays details of the three visually-grounded speech datasets used in this paper. When computing duration statistics, we exclude utterances longer than 15s for SpokenCOCO and Flickr8k Audio, and 40s for Places Audio, because we found that those utterances resulted from incorrect operation of the data collection interface (e.g., workers forgot to stop recording). When computing vocabulary sizes and word statistics, text transcripts are normalized by lower-casing all the alphabets and removing characters that are neither alphabets nor digits.

For the SpokenCOCO data collection on Amazon Mechanical Turk, we displayed the text of a MSCOCO caption to a user and asked them to record themselves reading the caption out loud. For quality control, we ran a speech recognition system in the background and estimated the word-level transcription for each recording. We computed the word error rate of the ASR output against the text that the user was prompted to read, and only accepted the caption if the word error rate was under 30%. In the case that the word error rate was higher, the user was asked to re-record their speech. We paid the users $0.015 per caption recorded, which in conjunction with the 20% overhead charged by Amazon resulted in a total collection cost of $10,898.91.

	SpokenCOCO	Flickr8k Audio	Places Audio
#Utts	605495	40000	400000
#Spks	2353	183	2683
#Imgs	123287	8000	400000
#Utts-per-img	5	5	1
Utt duration μ	4.12s	4.33s	8.37s
Utt duration σ	1.31s	1.33s	4.53s
#Words/utt	10.45	10.81	19.29
#Words/sec.	2.41	2.63	2.31
Duration	742hr	46hr	936hr
Vocab Size	19683	8718	41217
Type	scripted	scripted	unscritped

Table A1: Statistics and properties of the three visually-grounded speech datasets used in the paper.

B Detailed Experimental Setups

In this section, we provide details about data preprocessing, model architecture, and training hyperparameters for each module used in this paper. The same setups are used for all unit types unless otherwise stated.

B.1 Image-to-Unit Model

Data Images are reshaped to 256×256×3 matrices and are per-channel normalized with $\boldsymbol{\mu}$ = [0.485, 0.456, 0.406] and $\boldsymbol{\sigma}$ =[0.229, 0.224, 0.225]. During training, unit sequences are truncated or padded to the target length shown in Table A2. The target lengths are determined such that there are less than 10% sequences truncated while still allowing a reasonable batch size to be used. Units that occurred less than five times are excluded. Sequences are not truncated during evaluation. We follow the data splits used in (Harwath et al., 2020) for Places, and (Karpathy and Fei-Fei, 2015) for Flickr8k and SpokenCOCO (the "Karpathy split").

	Word	Char	VQ3	VQ2	WVQ	VQ3 \ RLE
Target Length	18	70	100	200	110	160
Sequence Truncated (%)	1.12	1.74	6.90	9.37	7.80	6.35
Batch Size (SAT)	80	60	40	40	40	40
Batch Size (SAT-FT)	32	32	20	-	-	-

Table A2: Configuration for each type of units used in the Image-to-Unit model.

Model We adopt an open-source re-implementation[2] of Show, Attend, and Tell (Xu et al., 2015) (SAT) with soft attention, which replaces the original CNN encoder with a ResNet-101 pre-trained on ImageNet for image classification. The last two layers of the ResNet are removed (a pooling and a fully-connected layer) such that the encoder produces a 14×14×2048 feature map for each image.

Training Adam (Kingma and Ba, 2015) with a learning rate of 10^{-4} is used for optimizing both stages (SAT and SAT-FT). The training objective is maximum likelihood combined with a doubly stochastic attention regularization introduced in (Xu et al., 2015) with a weight of 1. Dropout is applied to the input of decoder softmax layer with a probability of 0.5 during training. Gradients are clipped at 5 for each dimension. The first stage is trained for at most 30 epochs, and the best checkpoint from which is used to initialize the second

[2] Link to the SAT implementation on Github

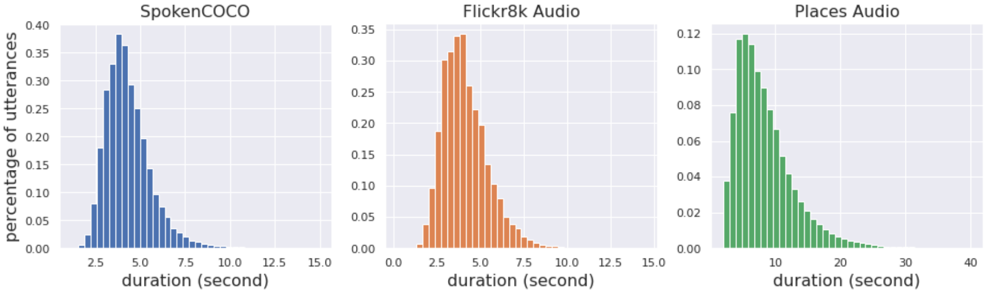


Figure A1: Utterance duration histograms for the three visually-grounded speech datasets.

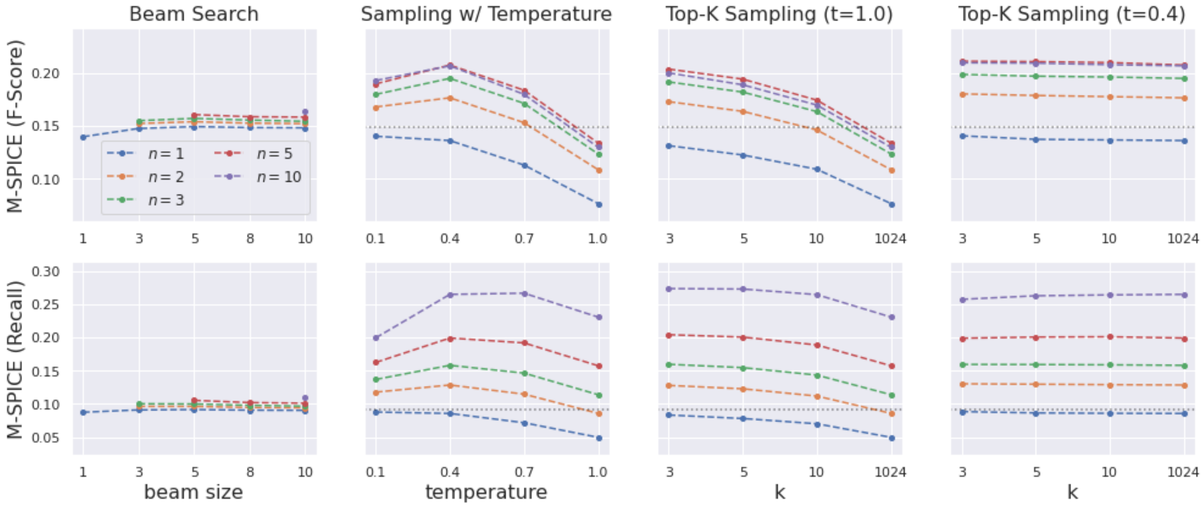


Figure A2: M-SPICE F-score (same as Figure 5) and recall on the SpokenCOCO test set with different candidate proposal methods.

stage trained for at most another 20 epochs. Models are selected based on the unit BLEU-4 score on the validation set. Using two NVIDIA TITAN X Pascal GPUs with data parallel training, each epoch takes about 2.8 hours for VQ3 units and 5.3 hours for VQ2 units.

B.2 Unit-to-Speech Model

Data RLE-ed unit sequences are used as input for all systems (VQ3 and VQ3 \ RLE systems share the same U2S model). The native audio sample rates in LJSpeech and VCTK are 22050Hz and 48kHz, respectively. For consistency and compatibility with the spectrogram-to-waveform model, we down-sample those in VCTK to 22050Hz. Following Tacotron2, we compute a 80 dimensional Mel spectrogram for each audio file with a 256-sample (11.6ms) frame hop, a 1024-sample (46.4ms) frame size, and a Hann window function. Utterances longer than 8 seconds are discarded during training to accommodate for the GPU memory constraints. We follow the data splits provided at `https://github.com/NVIDIA/tacotron2` for LJSpeech. For the multi-speaker VCTK dataset, we randomly sample 2.5% of the utterances from each speaker for validation.

Model We use an re-implementation[3] of Tacotron2 (Shen et al., 2018) for U2S models. For single-speaker models trained on LJSpeech, the exact same hyperparameters and model architecture are used as (Shen et al., 2018). For multi-speaker models trained on VCTK, we create an additional speaker embedding table of 256 dimensions for all speakers and control the speaker identity through these speaker embeddings. Speaker embeddings are injected at two places in the decoder: first in concatenation with the original input to the decoder LSTM, and second in concatenation with the output of the decoder LSTM, right before predicting the stop token and the spectra of a frame. A pre-trained[4] WaveGlow (Prenger et al., 2019) vocoder is used for all U2S models, which demonstrates the universality of vocoder models

[3] `https://github.com/NVIDIA/tacotron2`
[4] `https://github.com/NVIDIA/waveglow`

Metric	symbol	Sampling with Temperature				Top-K Sampling ($t = 1.0$)			Top-K Sampling ($t = 0.7$)		
		$t = 1.0$	$t = 0.7$	$t = 0.4$	$t = 0.1$	$k = 10$	$k = 5$	$k = 3$	$k = 10$	$k = 5$	$k = 3$
BLEU-4	VQ3	0.052	0.097	0.132	0.137	0.084	0.108	0.120	0.109	0.119	0.124
	VQ2	0.039	0.058	0.068	0.066	0.059	0.068	0.069	0.064	0.070	0.071
	WVQ	0.033	0.047	0.025	0.012	0.056	0.050	0.037	0.052	0.042	0.025
	VQ3 \ RLE	0.049	0.075	0.035	0.000	0.070	0.087	0.092	0.082	0.094	0.093
METEOR	VQ3	0.124	0.151	0.168	0.165	0.147	0.160	0.166	0.159	0.165	0.168
	VQ2	0.115	0.134	0.146	0.140	0.134	0.142	0.147	0.140	0.144	0.147
	WVQ	0.096	0.106	0.078	0.069	0.112	0.104	0.088	0.105	0.094	0.080
	VQ3 \ RLE	0.119	0.135	0.055	0.002	0.136	0.146	0.148	0.141	0.144	0.141
ROUGE-L	VQ3	0.303	0.358	0.403	0.416	0.346	0.371	0.386	0.373	0.386	0.397
	VQ2	0.293	0.330	0.351	0.345	0.325	0.345	0.351	0.340	0.348	0.355
	WVQ	0.270	0.297	0.287	0.287	0.312	0.309	0.292	0.309	0.295	0.276
	VQ3 \ RLE	0.295	0.330	0.152	0.001	0.328	0.349	0.355	0.340	0.348	0.350
CIDEr	VQ3	0.195	0.345	0.461	0.451	0.312	0.383	0.424	0.395	0.431	0.444
	VQ2	0.143	0.231	0.272	0.267	0.220	0.260	0.277	0.251	0.270	0.278
	WVQ	0.095	0.150	0.044	0.009	0.180	0.145	0.082	0.154	0.116	0.055
	VQ3 \ RLE	0.182	0.277	0.130	0.000	0.260	0.316	0.340	0.304	0.328	0.332
SPICE	VQ3	0.063	0.093	0.111	0.114	0.086	0.100	0.108	0.100	0.106	0.109
	VQ2	0.052	0.074	0.086	0.087	0.073	0.082	0.085	0.079	0.084	0.087
	WVQ	0.035	0.046	0.019	0.011	0.051	0.042	0.026	0.043	0.034	0.020
	VQ3 \ RLE	0.060	0.078	0.034	0.001	0.077	0.087	0.091	0.083	0.088	0.086

Table A3: Results of SAT models trained on MSCOCO and decoded with various sampling methods.

n	Beam Search					Sampling (t: temperature; k: top-k)									
	beam size=?					$(t, k) = (?, All)$				$(t, k) = (1.0, ?)$			$(t, k) = (0.7, ?)$		
	1	3	5	8	10	1.0	0.7	0.4	0.1	10	5	3	10	5	3
1	551	479	447	421	411	1447	978	689	561	1058	908	770	694	663	670
2	-	572	523	502	474	2100	1367	917	696	1522	1289	1025	907	867	851
3	-	693	620	585	562	2550	1644	1075	803	1855	1515	1222	1069	1003	973
5	-	-	681	625	617	3239	2111	1305	938	2367	1861	1511	1266	1209	1155
10	-	-	-	-	700	4311	2876	1664	1155	3176	2512	1954	1618	1552	1437

Table A4: The vocabulary size of the VQ3 SAT-FT model as estimated by various decoding approaches. The numbers in this table display the specific values of the curves depicted in Figure 4.

and how little acoustic properties of interest are affected by them.

Training A batch size of 64 are used for all systems. Adam (Kingma and Ba, 2015) with an initial learning rate of 10^{-3} is used to minimize the mean square error from spectrogram prediction and the binary cross entropy from stop token prediction combined. L2 regularization for the parameters with a weight of 10^{-6} is applied, and the L2 norm of the gradients are clipped at 1. Models are trained for 500 epochs on LJSpeech and 250 epochs on VCTK, and selected based on the validation loss. Empirically, each training epoch on LJSpeech takes about 12 minutes using two NVIDIA Titan X Pascal GPUs for both VQ2 and VQ3 models.

C Full Results of Decoding via Sampling

Table A3 presents the word-based evaluation results of decoding via sampling for all 5 metrics, supplementing Figure 3 in the main paper that only presents the SPICE results. We see that ranking between symbols are generally consistent among all those metrics, except the ranking between WVQ and VQ3 \ RLE when sampling with a temperature of 0.4. This is a relatively low-score regime when both model are transiting from generating trivial caption ($t = 0.1$) to non-trivial captions ($t = 0.7$).

D Full Results of Learned Vocabulary Size

In Table A4, we display the numerical results depicted graphically in Figure 4.

E More Image-to-Speech Samples

Table A5 shows captions sampled from the VQ3 model trained on MSCOCO. Here, we note that the sampled captions exhibit diversity both their content and linguistic style. We observe that the captioning model has learned to produce captions that correctly use quantifiers and conjugate verbs (“a couple of cows walking” vs. “a cow is standing”). The model also disentangles object identity from attributes such as color “red fire hydrant” vs. “yellow fire hydrant” vs. “green fire hydrant”).

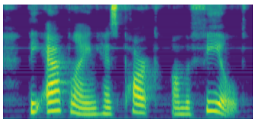
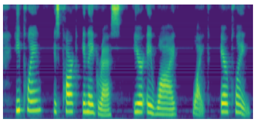
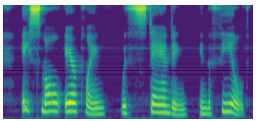
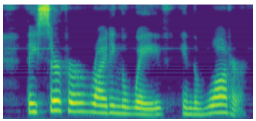
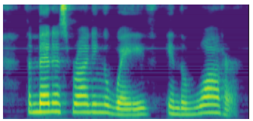
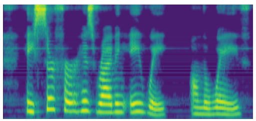
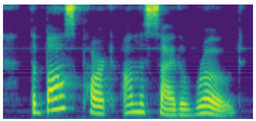
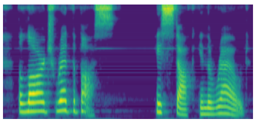
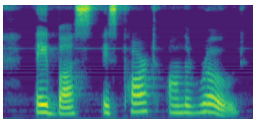
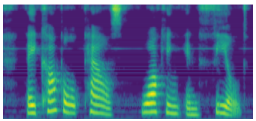
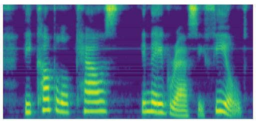
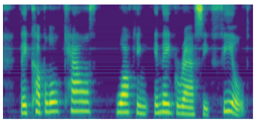
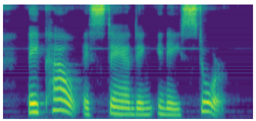
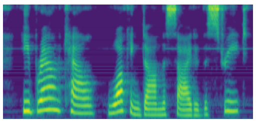
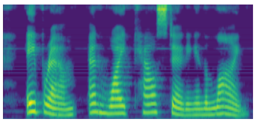
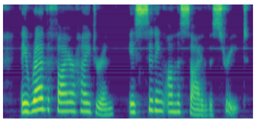
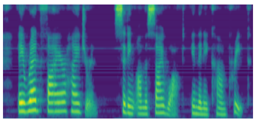
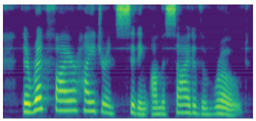
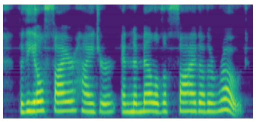
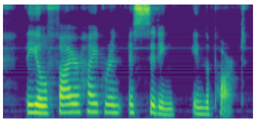
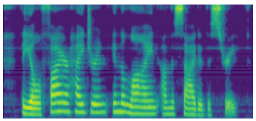
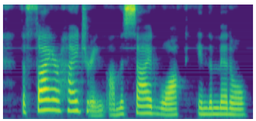
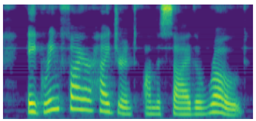
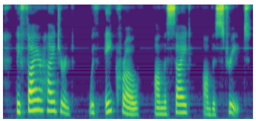
Image	Generated Spoken Captions / Transcripts (SAT-FT, VQ3, Sampling $(t, k) = (0.4, 3)$)		
	trial 1	trial 2	trial 3
	 the airplane is parked on the field	 a plane is parked in the grass near a white and white airplane	 a small airplane that is standing in a field
	 a surfer riding a wave in the water	 the man is riding the wave in the water	 a surfer is riding a wave on a wave
	 the bus parked on the side of the road	 a large red bus is stopped in the road	 a bus is parked on the road
	 a couple of cows walking in a field	 a couple of cows in a grassy field	 a couple of cows walking in a grassy field
	 a cow is standing in a store	 a brown cow walking down the side of a street	 a brown and white cow standing in a line
	 a red fire hydrant is sitting on the side of the street	 a red fire hydrant sitting on a sidewalk in a concrete	 a red fire hydrant sitting on the side of a road
	 a yellow fire hydrant in the middle of the side of a road	 a yellow fire hydrant is sitting in the park	 a yellow fire hydrant in a line on the side of a street
	 a fire hydrant on a sidewalk in the middle	 a green fire hydrant on the side of the road	 a fire hydrant with a curb on the side of the street

Table A5: Samples. More at https://wnhsu.github.io/image-to-speech-demo/2_vq3_sample_diversity_sat-ft_model