

---

# Text-Free Image-to-Speech Synthesis

## Using Learned Segmental Units – Appendix

---

<b>Wei-Ning Hsu</b> <sup>*†</sup> MIT wnhsu@csail.mit.edu	<b>David Harwath</b> <sup>*</sup> UT Austin harwath@utexas.edu	<b>Christopher Song</b> JHU csong23@mit.edu	<b>James Glass</b> MIT glass@mit.edu
---	--	---	--

### A Visually-Grounded Speech Datasets

Table A1 displays details of the three visually-grounded speech datasets used in this paper, and distributions of utterance duration are illustrated in Figure A1. 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.

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

	SpokenCOCO	Flickr8k Audio [5]	Places Audio [7]
Num. of Utterances	605495	40000	400000
Num. of Speakers	2353	183	2683
Num. of Images	123287	8000	400000
Num. of Utterances / Image	5	5	1
Utterance Duration $\mu$	4.12s	4.33s	8.37s
Utterance Duration $\sigma$	1.31s	1.33s	4.53s
Avg. Num. of Words / Utterance	10.45	10.81	19.29
Avg. Num. of Words / Second	2.41	2.63	2.31
Total Duration	742hr	46hr	936hr
Vocabulary Size	29539	8718	41217
Type	scripted	scripted	spontaneous

### B Detailed Explanation for M-SPICE

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

---

\*Equal contribution

†The author performed the work while at MIT, and is now at Facebook AI Research

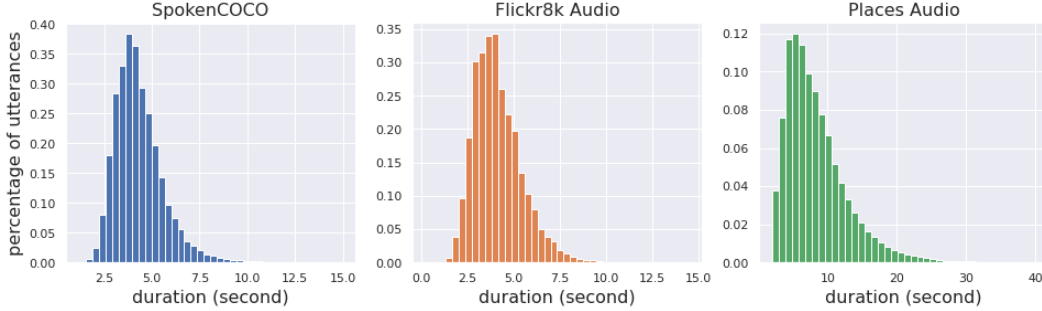


Figure A1: Utterance duration histograms for the three visually-grounded speech datasets.

from a scene graph parsed  $c$ , and we can compute that for multiple sentences with  $T(S) = \cup_i T(s_i)$ , and we have  $|T(S)| \geq |T(s_i)|$  for all  $i$ , as captions may capture different aspects of the same image.

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 [101]:  $max_j F1(T(S), T(c_j))$ . However, these metrics fail to capture the diversity among hypotheses. Let us now consider two hypothesis set,  $C^1 = \{c_1^1, c_2^1\}$  and  $C^2 = \{c_1^2, c_2^2\}$ , where

- $T(c_1^1) = T(c_2^1) = T(c_1^2) = \{(girl), (table), (girl, sit-at, table)\}$
- $T(c_2^2) = \{(girl), (girl, young)\}$
- $T(S) = \{(girl), (table), (girl, young), (girl, sit-at, table)\}$

We would like a metric to score  $C^2$  higher than  $C^1$  because it captures diverse and correct concepts; however, since F1 scores are the same for  $c_1^1, c_2^1, c_1^2$ , and is lower for  $c_2^2$ , the average SPICE of  $C^1$  is higher while the oracle SPICE are the same for both  $C^1$  and  $C^2$ . Our proposed M-SPICE can be formulated as  $F1(\cup_i T(s_i), \cup_j T(c_j))$ . When hypotheses capture diverse AND correct propositions, the M-SPICE recall should increase as shown in Figure A2 row 2. Note that the M-SPICE score and recall will not increase if propositions are diverse but incorrect.

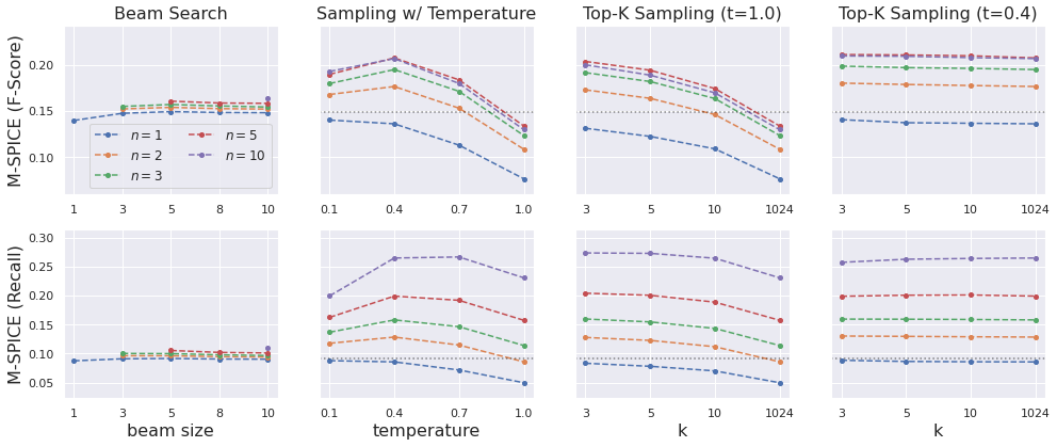


Figure A2: M-SPICE F-score (same as Figure 5) and recall on the SpokenCOCO test set with different candidate proposal methods.

## C Detailed Experimental Setups

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

## C.1 Image-to-Unit Model

**Data** Images are reshaped to  $256 \times 256 \times 3$  matrices and per-channel normalized with  $\mu = [0.485, 0.456, 0.406]$  and  $\sigma = [0.229, 0.224, 0.225]$ . 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 [6] for Places, and [10] for Flickr8k and SpokenCOCO (commonly known as the “Karpathy split”).

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

	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	-	-	-

**Model** We adopt an open-source re-implementation<sup>3</sup> of Show, Attend, and Tell [17] (SAT) with soft attention, which replaces the CNN encoder in [17] with a ResNet-101 [8] pre-trained on ImageNet [2] for image classification. The last two layers of the ResNet are removed (a pooling layer and a fully-connected layer) such that the encoder produces a  $14 \times 14 \times 2048$  feature map for each image.

**Training** Two model variants are considered in this paper: SAT and SAT-FT, which differ in how each part is initialized and which parts are updated during training. The SAT model initializes the encoder parameters with a pre-trained image classification model and freezes the encoder parameters during training. On the other hand, the SAT-FT model (fine-tuned SAT model) initializes the entire model with a pre-trained SAT model, and update all parameters during training. Adam [11] with a learning rate of  $10^{-4}$  is used for optimizing both models. The training objective is maximum likelihood combined with a doubly stochastic attention regularization introduced in [17] 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 batch size for each unit is shown in Table A2, which are chosen based on the target length and the GPU memory constraints. All SAT models are trained for at most 30 epochs, and SAT-FT models are trained for at most 20 epochs. Models are selected based on the unit BLEU-4 score on the validation set.

The time complexity of forward computation is the same for the encoder for all units, while for the decoder it is proportional to the unit sequence length due to the auto-regressive nature. 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.

## C.2 Unit-to-Speech Model

**Data** Run-length encoded unit sequences are used as input for all systems (i.e., VQ3 and VQ3 \ RLE systems share the same unit-to-speech model). The native sample rates of audio files in LJSpeech [9] and VCTK [15] are 22050Hz and 48kHz, respectively. For consistency and compatibility with the spectrogram-to-waveform model, we down-sample those in VCTK to 22050Hz. Following Tacotron [16] and Tacotron2 [14], we compute a 80 dimensional Mel spectrogram for each audio file with a 256-sample frame hop, a 1024-sample frame size, and a Hann window function. At a sample rate of 22050Hz, it corresponds to about 11.6ms frame hop and 46.4ms frame size. 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, to ensure the same speaker distribution between train and valid splits, we randomly sample 2.5% of the utterances from each speaker for validation, which results in a set of 1087 utterances.

<sup>3</sup><https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning>

**Model** We use a cascade of two systems to convert a unit sequence into a waveform. The first part synthesizes a Mel spectrogram from a sequence of units, which is referred to as the Unit-to-Speech model in this paper and determines most of the properties of interest in speech (e.g., linguistic content, speaker, prosody). The second part is a vocoder that converts a Mel spectrogram into a waveform, which mostly affects the fidelity of the synthesized speech rather than the aforementioned properties. A vocoder can be either a learned module like WaveNet [12] and WaveGlow [13], or a parameter-less signal processing block such as using the Griffin-Lim algorithm [4].

We use an re-implementation<sup>4</sup> of Tacotron2 [14] for Unit-to-Speech models, which is a sequence-to-sequence encoder-decoder model with location-sensitive attention [1]. For single-speaker models trained on the LJSpeech dataset, the exact same hyperparameters and model architecture are used as [14]. For multi-speaker models trained on the VCTK dataset, 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<sup>5</sup> WaveGlow [13] vocoder is used for all Unit-to-Speech models, which demonstrates the universality of vocoder models and how little acoustic properties of interest are affected by them. Although it is possible to achieve a even higher fidelity score through training or fine-tuning the WaveGlow model on the re-synthesized spectrograms, we did not attempt to experiment with that, since the focus of this paper is to demonstrate the capability of generating fluent spoken captions and controlling properties like speaker identity independently.

**Training** A batch size of 64 are used for all systems. Adam [11] 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.

The time complexity of forward computation at the encoder is proportional to the unit-sequence length because of the bi-directional LSTM layer in the encoder. On the other hand, the number of the decoding steps is proportional to the duration of the speech at the decoder, which is independent of the choice of input representation; however, at each decoding step, the number of encoder outputs the attention module attends to is proportional to length of the unit sequence. Empirically, each training epoch on LJSpeech takes about 12 minutes using two NVIDIA Titan X Pascal GPUs for both VQ2 and VQ3 models despite that VQ2 sequences are in average twice as long as VQ3 sequences, which shows that time complexity are dominated by other computations.

### C.3 Speech-to-Unit Model

We obtain the ResDAVEnet-VQ “{2} → {2, 3}” model and the WaveNet-VQ (PA) model reported in [6] from the authors. Both models learn discrete representations for speech and are used to transcribe speech into a sequence of units in this paper. We use these models to extract unit sequences for all datasets without fine-tuning, which examines the robustness of these Speech-to-Unit models when applied to datasets of different domains. Table A3 compares the three types of units used in this paper (VQ3, VQ2, WVQ) extracted from these two models. For self-containedness, the ABX error rate for each unit reported in [6] are also included. The ABX test evaluates the phone discriminability of the learned units on the ZeroSpeech 2019 English test set [3]. Note that (1) VQ3 and WVQ have the same unit rate before run-length encoding, (2) VQ2 achieves the lowest ABX error rate, (3) and all units have a lower ABX error rate before run-length encoding.

## D Image-to-Unit Samples

Table A4 and A5 display unit captions for the same image generated from Image-to-Unit models trained on different learned units. Captions in Table A4 are decoded with beam search (beam size=5), and those in Table A5 are sampled from the model distribution with top-k sampling ( $k = 5$ ).

<sup>4</sup><https://github.com/NVIDIA/tacotron2>

<sup>5</sup><https://github.com/NVIDIA/waveglow>

Table A3: Properties of the three types of units and the two speech-to-unit models.

	VQ3	VQ2	WVQ
Source Model	ResDAVEnet-VQ	ResDAVEnet-VQ	WaveNet-VQ
Training Data	Places Audio+Image	Places Audio+Image	Places Audio
Training Objective	Contrastive Loss	Contrastive Loss	Reconstruction Loss
RLE ABX Error Rate	15.68%	13.06%	25.23%
Pre-RLE ABX Error Rate	14.52%	12.51%	24.87%
Pre-RLE Unit Rate	40ms	20ms	40ms

Table A4 shows that Image-to-Unit models trained on WVQ and VQ3 units without run-length encoding (VQ3 \ RLE) fail to produce reasonable captions using beam search for *all* images. In fact, generated captions are almost always the same among different images for these two models. The WVQ model generates captions looping the same bi-gram until exceeding the maximum length, while the VQ3 \ RLE repeats the same unit. On the other hand, the model trained on VQ2 units can sometimes produce reasonable captions, but it exhibit the same behavior as the WVQ model when it fails as shown here.

On the contrary, Table A5 shows that all four models are capable of generating non-trivial captions without looping via sampling.<sup>6</sup> The observation here is consistent with the evaluation results on the transcribed spoken captions presented in Table A7 and A8 and Figure A3.

Table A4: Exemplar beam search decoding results from SAT Image-to-Unit models.

Symbol	Captioned Generated with Beam Search (beam size=5)
VQ3	263 32 208 5 336 100 717 803 256 803 815 144 120 144 654 936 48 417 272 417 362 766 825 284 614 156 341 135 769 5 208 32 208 5 336 815 144 815 494 181 467 417 870 395 683 141 250 543 820 587 181 913 1013 467 5 208 32 208 5 467 360 606 360 801 1009 398 847 89 100 869 254 1003 442 42 791 417 272 141 766 362 614 156 341 135 769 5 208 32
VQ2	71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791...
WVQ	181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232...
VQ3 \ RLE	263 32...

<sup>6</sup>The model might still generate trivial captions when sampling with a temperature close to zero or setting a very small  $k$  with top- $k$  sampling, in which case sampling is similar to greedy decoding.

Table A5: Exemplar sampling results with  $(t, k) = (1.0, 5)$  from SAT Image-to-Unit models.

Symbol	Captioned Generated with top-k sampling ( $k = 5$ )
VQ3	263 208 467 717 288 426 986 72 44 341 151 801 1022 27 320 426 288 66 570 683 351 313 910 820 543 820 230 100 852 329 852 288 502 706 427 110 451 297 938 457 426 100 852 329 852 791 993 522 993 374 502 288 936 48 263 208 32
VQ2	71 791 71 791 71 791 71 191 175 51 139 359 173 599 307 419 133 621 85 165 315 883 175 191 71 791 71 48 511 765 983 873 314 409 333 267 409 734 229 787 184 937 886 254 934 666 973 19 947 227 805 967 883 175 48 695 511 655 806 491 647 507 343 867 819 655 699 491 136 221 513 996 675 581 467 652 488 186 3 183 311 613 371 463 314 21 238 910 238 657 230 82 270 868 643 78 391 940 922 49 771 986 147 947 19 957 862 957 95 7 819 695 1011 159 831 589 966 827 753 891 162 253 269 219 13 501 977 302 241 157 691 723 695 175 191 71 791 71 791 71 791 71 48 1007
WVQ	181 232 181 232 181 232 181 232 181 232 181 225 124 232 181 232 225 232 181 225 124 225 232 181 252 169 211 147 89 67 156 155 189 110 53 246 225 89 52 21 5 216 155 225 25 47 41 223 225 181 166 57 185 82 25 225 124 149 214 93 28 195 65 1 23 109 246 223 141 47 41 223 181 232 82 231 188 169 147 89 225 181 225 124 181 124 5 216 53 246 181 225 137 52 5 159 225 181 225 46 155 246 232 181 232 225 232 181 225 52 30 5 216 166 225 124 225 181 225 5 4 46 225 181 25 137 52 159 155 225 181 225 108 155 246 225 108 155 225 232 181 25 89 221 70 197 232 181 225 214 28 214 225 181 232 244 220
VQ3 \ RLE	263 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 208 208 5 5 336 100 803 256 560 417 870 870 870 968 910 250 543 820 587 909 909 181 717 48 936 48 224 176 284 538 133 807 715 39 27 27 476 5 5 476 570 395 395 683 313 141 250 250 587 587 494 909 922 181 100 100 827 119 66 272 417 766 766 766 614 614 156 341 135 135 181 913 913 1013 5 208 208 208 208 208 208 5 5 5 476 320 96 96 651 538 133 766 766 825 740 913 1013 467 5 208 208 32 32 32 32 208 208 5 5 336 501 254 254 254 254 1003 442 852 362 825 740 639 639 587 543 543 975 320 320 284 284 228 844 844 622 622 846 654 654 846 336 263 208 32 32 32 32 32 32 32 32 32 32 32 32 32

## E Caption Evaluation on Unit Sequences

The spoken caption evaluations presented in Section 3.2 utilized an automatic speech recognizer to transcribe the generated captions into words so that they could be compared against reference text captions. In the case that a speech recognizer is not available, we can perform evaluation directly on the speech unit representations by using the unit model to transcribe the reference spoken captions. Table A6 displays BLEU-4, METEOR, ROUGE, and CIDEr scores evaluated directly on the various speech unit representations; we cannot compute SPICE scores directly on the speech units because SPICE requires a dependency tree for each caption. It is important to note that for a given evaluation metric, the scores across the models are not directly comparable because their unit spaces are different. We do note that the relative ranking among VQ3, VQ2, and Wavenet-VQ is consistent across BLEU-4, METEOR, and ROUGE, however, VQ3 \ RLE achieves abnormally high scores on these metrics despite producing trivial captions for all images as shown in Table A4. 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, which incorporates TF-IDF weighting that tends to de-emphasize these kinds of uninformative patterns. The fact that the CIDEr score is 0 for both the VQ3 \ RLE and Wavenet-VQ models indicates that in general the captions produced by these models are uninformative. We posit that word-level evaluation is always preferable for spoken caption generation, but in the case that this is not possible the CIDEr metric may be the best option.

Table A6: Unit-based caption evaluation on MSCOCO test set. The beam size  $\in \{3, 5, 10\}$  was chosen for each model to maximize the CIDEr score. Note that the scores between different units are not directly comparable, because they are computed based different types of units.

symbol	Greedy / Beam-Search (SAT Model)			
	Unit BLEU-4	Unit METEOR	Unit ROUGE	Unit CIDEr
VQ3	0.176 / 0.274	0.178 / 0.196	0.280 / 0.328	0.121 / 0.215
VQ2	0.172 / 0.141	0.132 / 0.108	0.178 / 0.157	0.027 / 0.020
WVQ	0.019 / 0.020	0.048 / 0.048	0.081 / 0.081	0.000 / 0.000
VQ3 \ RLE	0.163 / 0.163	0.168 / 0.168	0.218 / 0.218	0.000 / 0.000

## F Full Results of Caption Evaluation on Word Sequences

Table A7 and A8 and Figure A3 present the complete caption evaluation results on transcribed word sequences for all 5 metrics, supplementing Figure 3 in the main paper that only presents the SPICE results. Note that the transcripts for WVQ and VQ3 \ RLE beam search captions are not reliable; for the reasons discussed in the previous section the generated spoken captions contain only silence, and the ASR model used for transcription did not see utterances comprised of pure silence during training. 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 ( $t = 0.7$ ).

Table A7: Word-based caption evaluation on MSCOCO test set. An ASR model is used to transcribe the spoken captions into text for evaluation. The beam size  $\in \{3, 5, 10\}$  was chosen for each model to maximize the SPICE score.

symbol	Greedy / Beam-Search (SAT Model)				
	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
word	0.287 / 0.315	0.247 / 0.253	0.524 / 0.533	0.939 / 0.984	0.180 / 0.185
char	0.238 / 0.289	0.230 / 0.239	0.495 / 0.512	0.783 / 0.879	0.164 / 0.172
VQ3	0.133 / 0.186	0.162 / 0.186	0.413 / 0.446	0.435 / 0.584	0.111 / 0.127
VQ2	0.068 / 0.073	0.138 / 0.126	0.343 / 0.345	0.262 / 0.224	0.084 / 0.065
WVQ	0.010 / 0.009	0.069 / 0.069	0.286 / 0.285	0.009 / 0.009	0.011 / 0.011
VQ3 \ RLE	0.000 / 0.000	0.002 / 0.002	0.001 / 0.001	0.000 / 0.000	0.001 / 0.001

Table A8: Sampling-based evaluations with SAT models trained on different units.

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



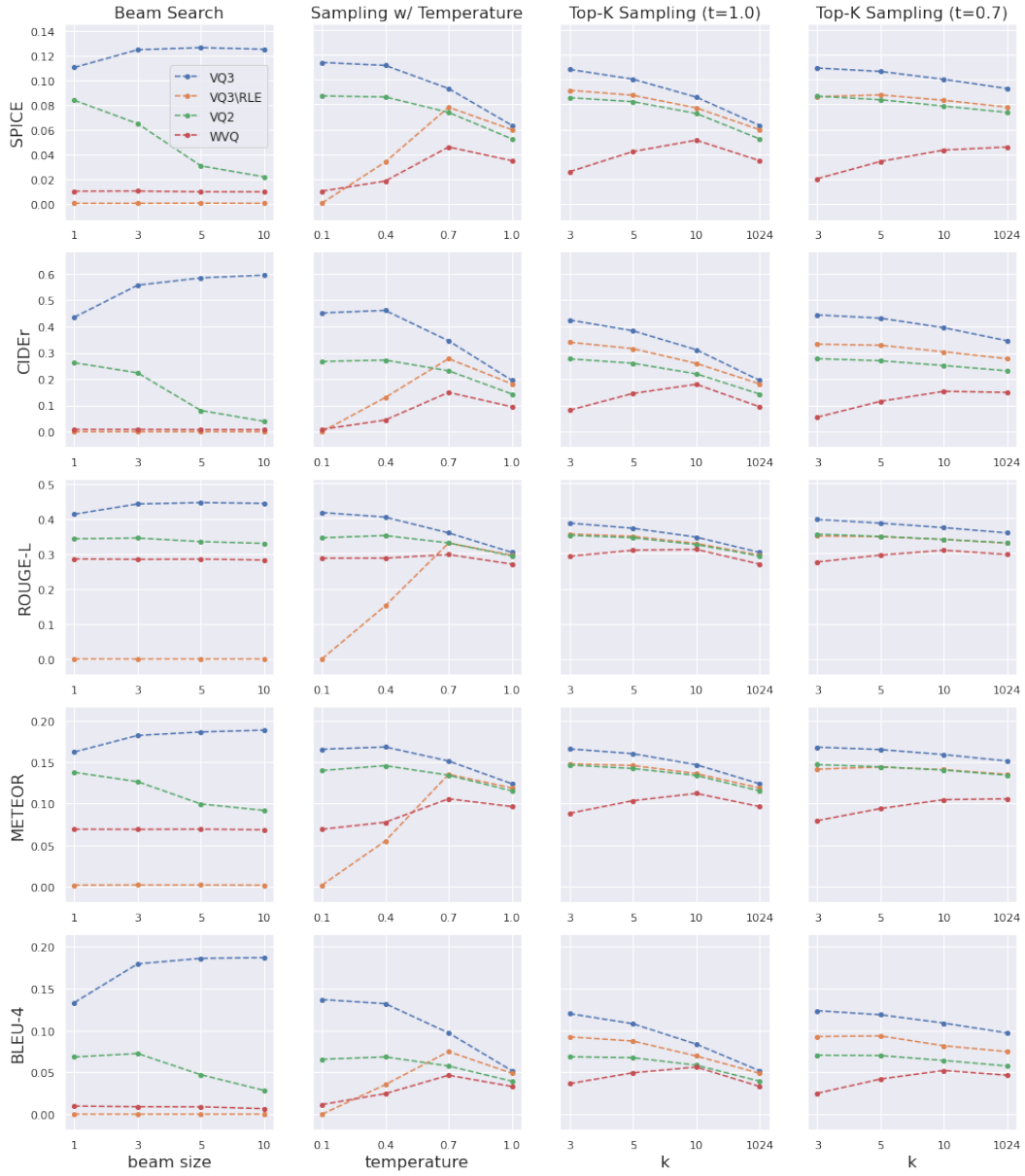


Figure A3: Word-based caption evaluation results of all five metrics on SAT models trained on different units, using both beam search decoding and sampling for unit sequence generation.

## G Comparison of Average SPICE and M-SPICE

Figures A4 and A5 display the M-SPICE scores and SPICE score distributions of different sampling methods for the SAT and SAT-FT model trained on VQ3 units, respectively. The exact numbers are shown in Tables A9, A11, A10, and A12. When performing sampling-based evaluation, there is bound to be some stochasticity in the results. However, the SPICE score distributions (box plots over 10 trials) shown in the bottom row of Figures A4 and A5 are very narrow, which we attribute to the fact that the COCO test set is large enough attenuate this stochasticity. The narrowness of the box plots also suggests that taking the average SPICE score over multiple sampling runs does not reflect the diversity of the captions the way that M-SPICE does.

Table A9: M-SPICE F1-scores of the VQ3 SAT model with beam search decoding.

$n =$	Beam Size				
	1	3	5	8	10
1	0.111	0.125	0.127	0.126	0.125
2	-	0.127	0.130	0.129	0.128
3	-	0.129	0.131	0.131	0.130
5	-	-	0.134	0.133	0.132
10	-	-	-	-	0.135

Table A10: M-SPICE F1-scores of the VQ3 SAT-FT model with beam search decoding.

$n =$	Beam Size				
	1	3	5	8	10
1	0.140	0.147	0.149	0.148	0.148
2	-	0.152	0.154	0.152	0.152
3	-	0.155	0.157	0.155	0.154
5	-	-	0.161	0.159	0.158
10	-	-	-	-	0.164

Table A11: M-SPICE F1-scores of the VQ3 SAT model with sampling.  $t$  denotes the temperature, and  $k$  denotes the number of top units considered at each decoding step for top-K sampling.

$n =$	Sampling with Temperature				Top-K Sampling ( $t = 1.0$ )			Top-K Sampling ( $t = 0.4$ )		
	$t = 1.0$	$t = 0.7$	$t = 0.4$	$t = 0.1$	$k = 10$	$k = 5$	$k = 3$	$k = 10$	$k = 5$	$k = 3$
1	0.063	0.093	0.111	0.114	0.086	0.100	0.108	0.100	0.106	0.109
2	0.092	0.128	0.147	0.137	0.120	0.137	0.145	0.138	0.143	0.146
3	0.106	0.145	0.163	0.147	0.137	0.153	0.161	0.153	0.160	0.163
5	0.117	0.156	0.175	0.154	0.148	0.165	0.173	0.164	0.173	0.174
10	0.115	0.153	0.173	0.155	0.145	0.160	0.169	0.162	0.169	0.171

Table A12: M-SPICE F1-scores of the VQ3 SAT-FT model with sampling.

$n =$	Sampling with Temperature				Top-K Sampling ( $t = 1.0$ )			Top-K Sampling ( $t = 0.4$ )		
	$t = 1.0$	$t = 0.7$	$t = 0.4$	$t = 0.1$	$k = 10$	$k = 5$	$k = 3$	$k = 10$	$k = 5$	$k = 3$
1	0.076	0.113	0.136	0.140	0.109	0.122	0.131	0.137	0.137	0.141
2	0.108	0.153	0.177	0.168	0.146	0.164	0.173	0.178	0.179	0.180
3	0.123	0.171	0.195	0.180	0.163	0.182	0.192	0.196	0.197	0.199
5	0.134	0.184	0.208	0.190	0.174	0.194	0.204	0.210	0.211	0.211
10	0.130	0.180	0.207	0.193	0.170	0.189	0.200	0.208	0.209	0.210

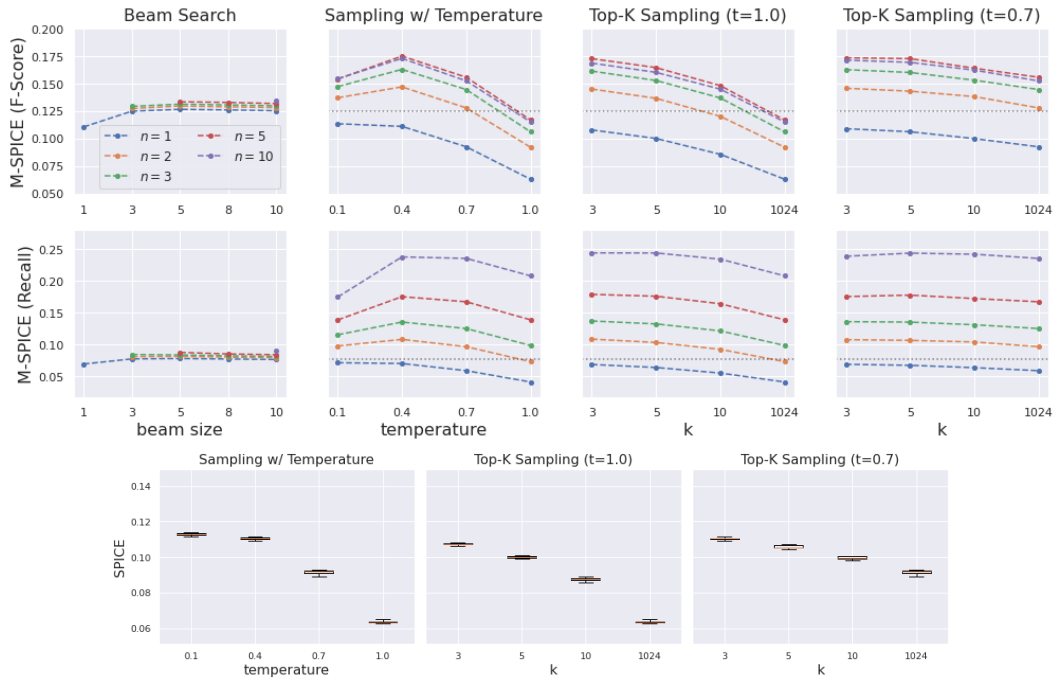


Figure A4: M-SPICE F1-scores and recalls (top) and SPICE distributions (bottom) of the SAT model on the MSCOCO test set with different caption generation methods. Box-and-whisker plots the SPICE scores over 10 runs are shown, where a box extends from the first quartile to the third quartile.

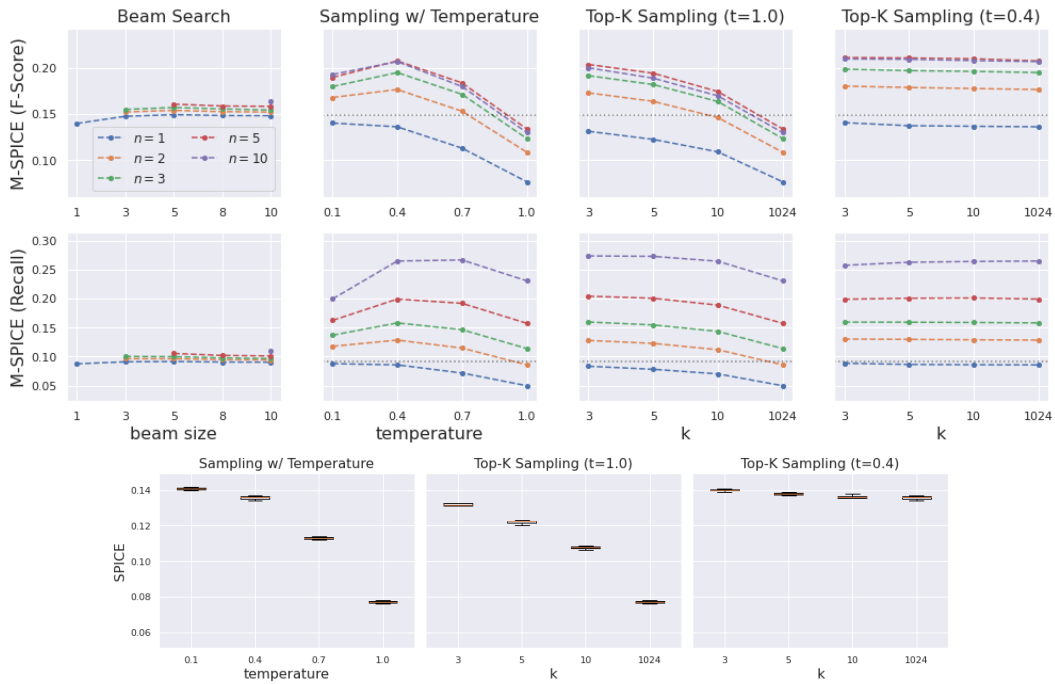


Figure A5: M-SPICE F1-scores and recalls (top) and SPICE distributions (bottom) of the SAT-FT model on the MSCOCO test set with different caption generation methods.

## H Full Results of Learned Vocabulary Size

In Table A13, we display the numerical results depicted graphically in Figure 4.

Table A13: 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.

$n$	Beam Search beam size=?					Sampling ( $t$ : temperature; $k$ : top-k)									
	1	3	5	8	10	$(t, k) = (?, All)$				$(t, k) = (1.0, ?)$			$(t, k) = (0.7, ?)$		
						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

## I 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 A14. In order to compute these scores we transcribe the captions generated by each model into text using the ASR system we describe in Section ??, which was solely trained on re-synthesized SpokenCOCO captions using the LJSpeech U2S model. The scores in Table A14 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. In Figure ??, we show example captions generated by conditioning on the same unit sequence, but different speaker identities.

Table A14: Demonstration of disentangled voice control via synthesizing the same units with different unit-to-speech models conditioned on different speaker IDs. Units are generated with beam search using the SAT-FT model for the MSCOCO test set.

Train Data	Speaker ID	Gender	Region	BLEU-4	METEOR	ROUGE	CIDER	SPICE
LJSpeech	-	F	-	0.233	0.212	0.478	0.732	0.149
VCTK	p247	M	Scottish	0.234	0.211	0.480	0.730	0.148
	p231	F	English	0.233	0.210	0.478	0.724	0.146
	p294	F	American	0.236	0.212	0.482	0.732	0.148
	p345	M	American	0.234	0.209	0.477	0.717	0.144
	p307	F	Canadian	0.234	0.211	0.479	0.729	0.148

## J More Image-to-Speech Samples

In Tables A15 and A16, we show many more examples of spoken captions generated by the VQ3 model. In Table A15, all three captions in each row were generated from the same unit sequence corresponding to the top hypothesis from beam search decoding. Each column represents the caption waveform generated by a different U2S model reflecting a different speaker. Although the spectrograms are visibly very different (reflecting the differing speaker characteristics across the U2S models), the word sequence estimated by the ASR model is generally the same. We do notice some substitution errors (highlighted in red), most commonly between “a” and “the”.

Table A16 shows captions generated by the same VQ3 model and for the same set of images depicted in Table A15, but instead of varying the U2S model we show captions generated via sampling rather than beam search. 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”).

Table A15: Samples. More at [https://xyn7a5e0vs.github.io/image-to-speech/3\\_vq3\\_voice\\_control\\_sat-ft\\_model](https://xyn7a5e0vs.github.io/image-to-speech/3_vq3_voice_control_sat-ft_model)


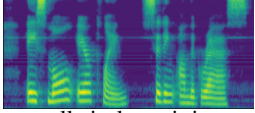
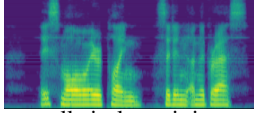
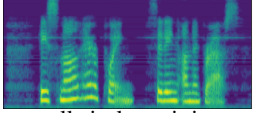

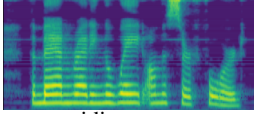
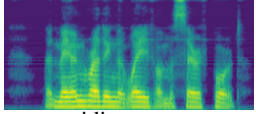
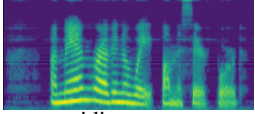

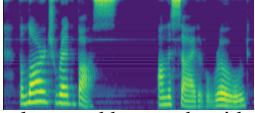
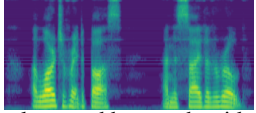
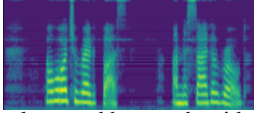

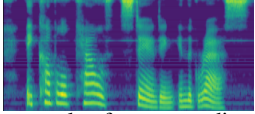
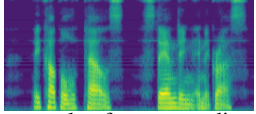
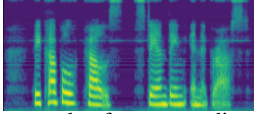

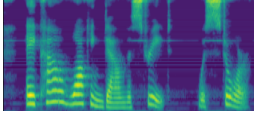
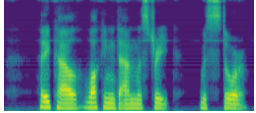
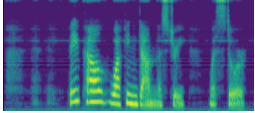

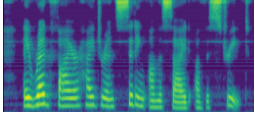
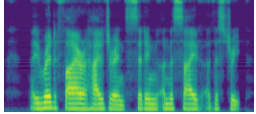
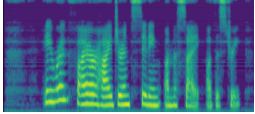

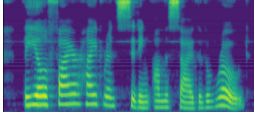
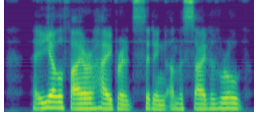
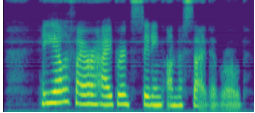

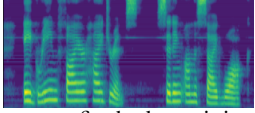
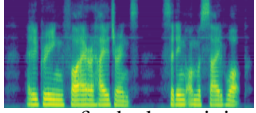
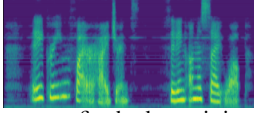
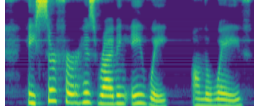

Image	Generated Spoken Captions / Transcripts (SAT-FT, VQ3, Beam Search)		
	LJSpeech	VCTK (p247)	VCTK (p307)
	 a small airplane sitting on the grass	 a small airplane sitting on the grass	 a small airplane sitting on the grass
	 a man riding a wave on a surfboard	 a man riding a wave on a surfboard	 a man riding a wave on a surfboard
	 a large red bus on the side of <b>the</b> road	 a large red bus on the side of <b>the</b> road	 a large red bus on the side of <b>a</b> road
	 a couple of cows standing in the grass	 a couple of cows standing in the grass	 a couple of cows standing in the grass
	 a cow walking down the street <b>with</b> a store	 a cow walking down the street <b>in</b> a store	 a cow walking down the street <b>next to</b> a store
	 a red fire hydrant sitting on the side of a street	 a red fire hydrant sitting on the side of a street	 a red fire hydrant sitting on the side of a street
	 a yellow fire hydrant sitting on the side of <b>a</b> road	 a yellow fire hydrant sitting on the side of <b>the</b> road	 a yellow fire hydrant sitting on the side of <b>a</b> road
	 a green fire hydrant sitting on <b>a</b> sidewalk	 a green fire hydrant sitting on <b>the</b> sidewalk	 a green fire hydrant sitting on <b>a</b> sidewalk

Table A16: Samples. More at [https://xyn7a5e0vs.github.io/image-to-speech/2\\_vq3\\_sample\\_diversity\\_sat-ft\\_model](https://xyn7a5e0vs.github.io/image-to-speech/2_vq3_sample_diversity_sat-ft_model)

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

## References

- [1] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2015.
- [2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
- [3] 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.
- [4] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing (TASSP)*, 32(2):236–243, 1984.
- [5] 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.
- [6] David Harwath, Wei-Ning Hsu, and James Glass. Learning hierarchical discrete linguistic units from visually-grounded speech. In *Proc. International Conference on Learning Representations (ICLR)*, 2020.
- [7] 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.
- [8] 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.
- [9] Keith Ito. The LJ speech dataset. <https://keithito.com/LJ-Speech-Dataset/>, 2017.
- [10] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations (ICLR)*, 2015.
- [12] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.
- [13] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
- [14] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 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, 2018.
- [15] Christophe Veaux, Junichi Yamagishi, and Kirsten Macdonald. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit. 2017.
- [16] 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. Tacotron: Towards end-to-end speech synthesis. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2017.
- [17] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. International Conference on Machine Learning (ICML)*, 2015.