Text-Free Audio Captions of Short Videos from Latent Space Representation

by

Anisha Agarwal

S.B. in Computer Science and Engineering
Massachusetts Institute of Technology (2021)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

Author .....
Department of Electrical Engineering and Computer Science
May 13, 2022

Certified by .....
Aude Oliva
Principal Investigator
Thesis Supervisor

Accepted by .....
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Text-Free Audio Captions of Short Videos from Latent Space Representation

by

Anisha Agarwal

Submitted to the Department of Electrical Engineering and Computer Science
on May 13, 2022, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

In this thesis, we re-implement previous work exploring image to speech captioning. We expand upon the work to implement video to speech captioning. Specifically, we implement a text-free image to speech captioning pipeline that integrates four distinct machine learning models. We alter the models to process video data rather than image data and analyze the resulting speech captions. We conduct experiments on the Wav2Vec2 and HuBERT Automatic Speech Recognition models, and identify which works best with synthesized speech.

Thesis Supervisor: Aude Oliva
Title: Principal Investigator

Acknowledgments

I've learned so much during my time with the Computational Perception and Cognition Lab. Thank you to Aude Oliva for creating such a warm and intellectually creative environment, and thank you to everyone for welcoming me into the group.

Thank you especially to SouYoung Jin for her tireless guidance and assistance throughout the year.

Thank you to my parents and brother for their continued support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The ability to automatically generate dense captions and summaries of videos has an astounding number of real world applications. It can be used to provide people with visual or audio processing disabilities greater access, or create summaries of long videos so viewers can quickly gain an understanding of long-form media. As valuable as the ability to generate dense *written* captions is the ability to generate dense *spoken* captions; human writing and speech are distinct from one another, and to map machine writing and machine speech to the same problem space would be a vast oversight. We aim to outline and begin work on the first model that proceeds directly from videos to spoken captions without text as an in-between representation at any point in the pipeline.

In this paper, we suggest a pathway for automatically generating spoken captions from videos. First, we implement a model that takes images as input and outputs spoken captions. This model is a re-implementation of the work done by Hsu et al [12], which uses a latent space representation of 1024 phonic units (also referred to as phenomes) rather than a vocabulary of specific words. A phenome-based vocabulary is more lightweight and generalizable across disparate domains. Such a vocabulary also has the ability to represent information in a manner that is more in line with the way people speak, rather than the way people write. Their pipeline combines four machine learning models: the ResDAVEnet-VQ model [8], the Show, Attend, Tell

model (SAT) [39], the Tacotron2 model [35] and the WaveGlow model [30].

Second, we alter and re-train the model to take short videos as input rather than singular images. We do so by creating an input with an additional dimension representing the number of frames in our video. The latent space phonic unit representation does not change with this alteration; we simply provide the model with more information and train it to do the same task of captioning the input.

The third step of this process is to adjust the architecture even further to handle untrimmed video rather than short videos. That is outside the scope of this project, but would be valuable as future work.

1.1 Dense Video Captioning

Our long-term goal is to create a model with the ability to generate dense spoken descriptions from untrimmed videos. The current state-of-the-art models for dense video captioning, such as [42, 23], focus on generating written captions rather than spoken captions. While such work provides promising results for dense video captioning, we begin work on a new pathway towards the same end goal. We investigate latent space phonic representations and audio captions rather than text captions, thus broadening the horizon of this active field of study within machine learning.

1.2 Text Free Image to Speech Captioning

The majority of this work is focused on re-implementing, expanding and experimenting upon the baseline architecture for Text-Free Image to Speech Captioning outlined in [12]. We do so because this paper is the first to attempt image to speech captions without explicit text as an in-between representation. Using phenome units rather than words has many implications in the model training process: a vocabulary of 1024 units could replace larger and less generalizable word-based vocabularies.

We believe this nuance of their implementation can potentially shift the domain of outputs towards more conversational-sounding captions.

1.3 ASR Evaluation

As part of our evaluation pipeline, we use an Automatic Speech Recognition (ASR) model to transcribe synthesized speech and calculate often-used NLP scoring metrics on the resulting caption. This follows the procedure laid out by Hsu et al [12]. There, however, the authors train an entire ASR model to transcribe text synthesized by the speech generation module. While this is certainly the ideal, doing so requires Amazon Mechanical Turk (AMT) workers to manually transcribe the synthesized audio for ground truth data. As such, this ASR procedure becomes an expensive and prohibitive method to evaluate the performance of a model.

To mitigate this barrier, we experiment with open-source, free ASR models and assess the transcriptions from each. We intend to identify the open source model that best transcribes synthesized text, so that evaluation of the text-free pipeline is not halted by an inability to meaningfully evaluate the model's performance. We hope to provide researchers a less expensive, more rapid alternative for model evaluation.

Chapter 2

Related Work

2.1 Image to Text

In recent years, there have been significant improvements in text caption generation for images. This is predominantly driven by the creation of deep learning networks. Works such as [39], [15], and [2] demonstrate how effective machine learning models have become over recent years at the captioning task. In all three papers, researchers use deep learning networks to go directly from image input to written text output. Since the current state-of-the-art in this domain is so effective, further advancements and improvements in similar, but less explored, tasks have become even more prevalent.

2.2 Video to Text

In the video to text problem formulation, researchers build models that take as input entire videos, and output written captions for each video. Some models, such as [42] and [23] elect to output dense captions based on a sequence of temporal proposals across the input video.

Other work, such as [1], aims to skim long videos and output a brief summary of the events across the video. Here, the problem shifts from localizing many actions and captioning all of them to summarizing the video as concisely and accurately as

possible.

2.3 Text to Speech

Text to speech (TTS) is a well studied problem that involves generating an audio waveform to speak the words provided in written text. With the development of deep learning, neural network-based TTS has significantly improved the quality of synthesized speech. As highlighted in the high-level review done by Tan et al [36], models such as DeepVoice 1 and 2 [4, 3] improve the statistical synthesis of early TTS models with neural network-based additions. Following work to combine statistical and neural synthesis came WaveNet, which generates a waveform from linguistic features, and is considered to be the first fully neural TTS model [24]. The next steps in the field progressed from linguistic features as input to characters and phenomes as input [41, 35, 29, 32, 31]. The current state of the art in TTS involves end-to-end TTS systems that can generate waveforms directly from text, such as ClariNet [28], FastSpeech 2 [31] and EATS [7].

2.4 Speech to Text

Similarly to image captioning, Automatic Speech Recognition (ASR) has improved massively in recent years. In early years of study, Speech to Text models utilized a Hidden Markov Model-Gaussian Mixed Model to recognize speech. This architecture was discovered and popularized in [17]. This architecture was then improved with HMM-deep neural network (DNN) models and end-to-end models using deep learning, which have been shown to perform better than the original HMM-GMM baseline. This includes models such as [6, 5, 11].

2.5 Image/Video to Speech

2.5.1 Image to Speech

Many prior works have explored image caption generation, but very few have explored captioning images with spoken captions without an intermediate text representation of the caption. Prior to 2020, Hasegawa et al [9] was the only such example. In 2020, however, the work done by Hsu et al [12] explored and made enormous strides in automatically captioning images without a text-based latent space representation. They did so by representing language with phonic units rather than a vocabulary of words. They aimed to investigate, among other things, how written and spoken language differ. Recently, work in this problem space has expanded to the creation of datasets for an Image to Speech pipeline with minimized bias [25].

2.5.2 Video to Speech

Much of the work done to build models that take as input video and as output speech are motivated by helping people with visual impairments understand video content via Audio Descriptions (AD) [34, 40, 16]. Although such model pipelines do go from video to speech, they are not focused on the task of bypassing a textual representation.

In the Spoken Moments paper [21], researchers build a dataset mapping videos to spoken descriptions. Unlike previous such datasets with spoken captions, the captions are not generated by having annotators read written text out loud. Instead, annotators speak a description of the given video directly, and an ASR model transcribes their spoken caption. Since our end goal is to generate spoken captions that are different from written captions because of their closeness to speech, the Spoken Moments dataset is extremely useful.

Chapter 3

Methods

3.1 Building the Baseline Model

The brunt of the work described in this paper is centered around the model architecture outlined in [12]. Here, they introduce the idea of describing images with spoken captions without an intermediate step involving text. They replace text with units: 1024 integers that correspond to latent space representations of phonics rather than words explicitly. This larger problem can be broken down into 3 distinct problems as follows:

1. Gathering ground-truth data mapping Speech to Unit sequences (Speech to Unit)

2. Training a model to take as input an Image and output Units (Image to Unit)

3. Training a model to take as input Units and output a Speech audio file (Unit to Speech)

We can then string these 3 distinct models together to create a complete Image to Speech (I2S) pipeline that takes as input an image, uses units as an intermediate representation, and outputs a speech caption. The full diagram is drawn in Figure 3-1.

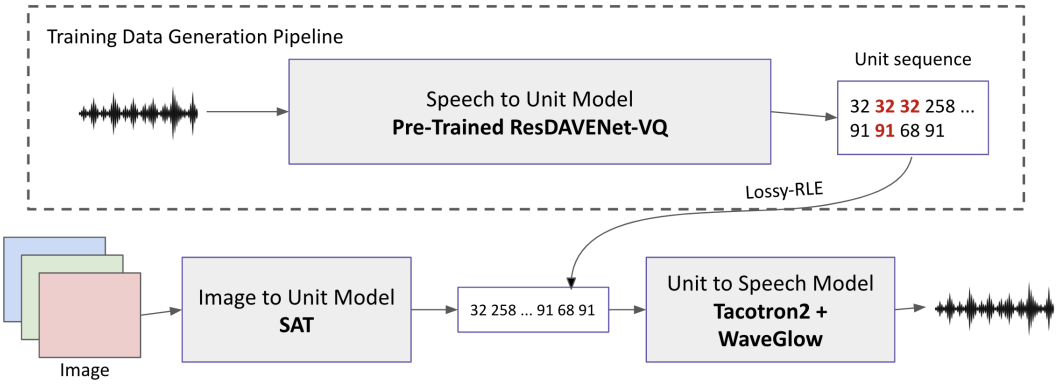


Figure 3-1: Full Image to Speech Pipeline, as defined by Hsu et al [12]

Note that we do not train the Image to Unit and Unit to Speech models together. Rather, we train them separately and combine the two only for full-pipeline inference.

3.1.1 Gathering Ground-Truth Data

We use ResDAVEnet-VQ as the Speech to Unit (S2U) model to gather the ground-truth data for our pipeline [8]. The relevant function of this model is its ability to extract quantitative features from audio input. By extracting the one-hot vector output from the third vector quantized (VQ) layer in the ResDAVEnet-VQ model, we get our desired sequence of phenome units representing the input audio. After gathering the unit sequences, we utilize 2 additional pre-processing steps before training models on this data:

1. Remove all units from the output sequences that are used less than 5 times across our data.

2. Run a lossy run-length-encoding on the resulting sequences, which finds subsections of unit sequences that consist of the same unit over and over, and removes all but the first instance of said unit. So, a sequence of units such as {238, 32, 32, 32, 71, 89, 71} becomes {238, 32, 71, 89, 71}.

We use 2 separate datasets and protocols for training the Image to Unit and Unit to Speech models. For the Image to Unit model, we train on the SpokenCOCO

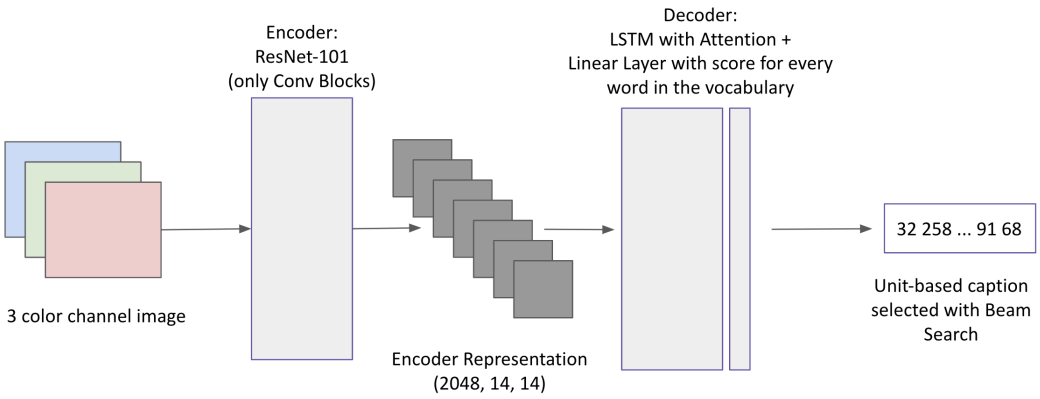


Figure 3-2: Image to Unit SAT Model architecture

dataset. Image inputs come directly from the COCO dataset [20], while the units are generated by passing the spoken version of captions from the COCO dataset through the ResDAVEnet-VQ model. We save the unit sequence output for each audio file in a dataset into its own text file, and read it into code during the data pre-processing step of downstream models.

For the Unit to Speech model, we use the LJSpeech dataset [13]. This dataset does not contain images, but rather is focused on text to speech synthesis. We pass the ground truth LJSpeech audio files through the ResDAVEnet-VQ model to attain ground truth units, and train the Unit to Speech model from those units to the unchanged ground truth speech files.

3.1.2 Image to Unit Model

The Image to Unit (I2U) model takes as input an image, and outputs a caption for said input. However, rather than a text caption, the I2U module outputs a sequence of units. Units are a lower-level representation of phonics. So, rather than being limited to a certain vocabulary to describe any image it could possibly see, the I2U module has phonic units to represent sounds, not words. Units are integers ranging from 0 to 1023. This means that the I2U module can ostensibly describe a wider range of images with a smaller vocabulary.

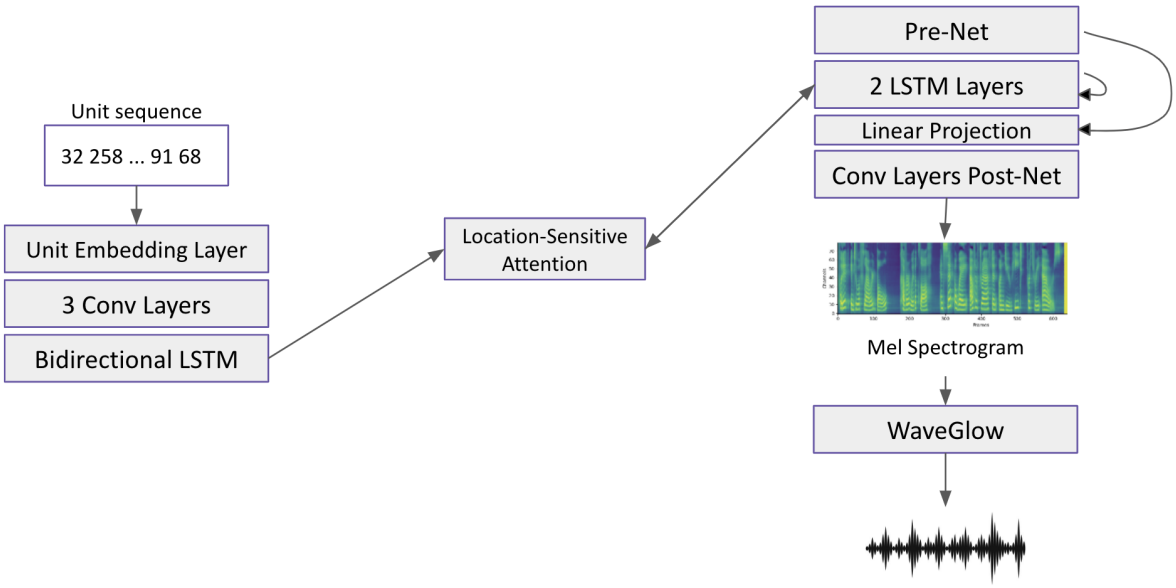


Figure 3-3: Unit to Speech Tacotron2 + WaveGlow Model Architecture

As outlined in Text-Free Image-to-Speech Synthesis Using Learned Segmental Units [12], we utilize an open-source implementation of a Show, Attend and Tell (SAT) image captioning model [39]. We alter the open-source implementation to output unit sequences as captions rather than written text. The SAT model (as seen in Figure 3-2) consists of an Encoder-Decoder architecture. The encoder is made up of the convolutional blocks of an off-the-shelf ResNet101 [10], pre-trained to encode images for downstream classification. The encoder takes as input an image and outputs an encoding of size 2048 x 14 x 14. The decoder outputs a sequence of units given the encoded image using an LSTM with attention [37]. We set the maximum length of a unit sequence to be 100 units, which truncates about 7% of the unit sequences after pre-processing. In our training, we train the decoder from scratch for 30 epochs, but we neither fine-tune nor train the encoder.

3.1.3 Unit to Speech Model

The Unit to Speech (U2S) model takes as input a sequence of units, and outputs a synthesized audio file based on the units. For our implementation, as per the method described by Hsu et al [12], we used the Tacotron2 model from NVIDIA, in tandem with the WaveGlow model [35, 30], as seen in Figure 3-3. The Tacotron2 model also employs an Encoder-Decoder structure with attention to perform the text-to-speech action. We alter the code from the open-source Tacotron2 repository to take as input unit sequences rather than text. The Tacotron2 model outputs a learned Mel spectrogram, and the training loss is computed by comparing the Mel spectrogram of the ground truth audio to the one of the synthesized audio. We train the Tacotron2 model for 500 epochs. Since unit sequences tend to be longer than the sentences of the training corpus (100 units vs no more than 50 words), training took a bit more time. In fact, the model was so large that it was impossible to run training at all with unit sequences of length 300 (which we tried before cutting the maximum length at 100).

During inference, a pre-trained WaveGlow model takes as input the Mel spectrogram from the Tacotron2 model and outputs an audio file. WaveGlow performs very well at Mel spectrogram comprehension across different speakers and contexts, so no fine-tuning or training was required.

The Tacotron2 model was trained with the LJSpeech dataset [13]. Since LJSpeech is an audio dataset of non-fiction read out loud by one person, it tends to have longer, more complex sentences and words than the SpokenCOCO dataset. The Tacotron2 model quantitatively shifts between different audio dataset domains very well: it manages to output viable, intelligible audio captions for unit sequence inputs during inferences even on input data outside of LJSpeech.

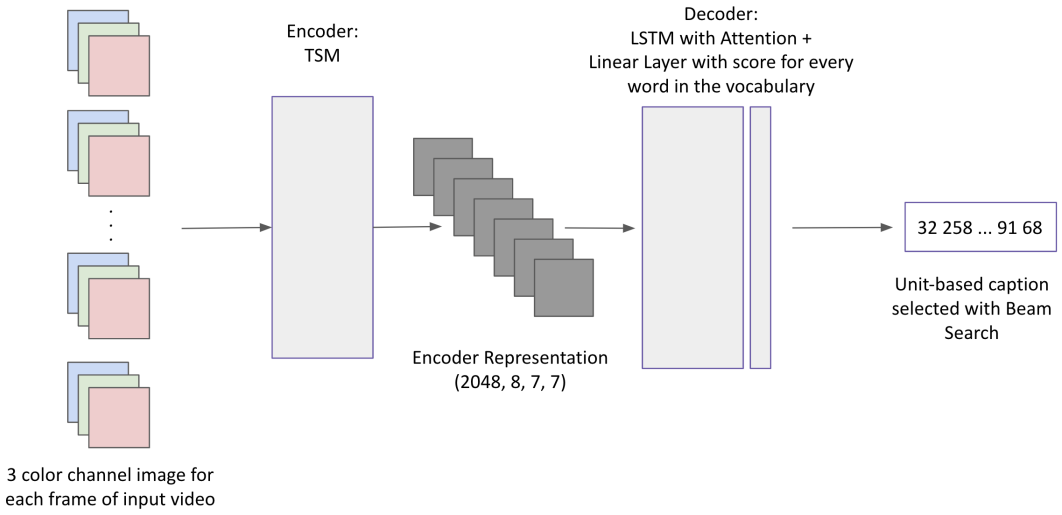


Figure 3-4: SAT Model Architecture, altered to take video frames as input instead of a single image.

3.2 Video to Speech Model

After following the methodology outlined by Hsu et al [12] to recreate their pipeline for generating audio captions for images, we alter the SAT model to accept video as input rather than an image, as seen in Figure 3-4. We give the Video to Speech (V2S) model a sequence of images representing each frame of the input video. We alter the original SAT model architecture by replacing the encoder from the SAT model with a Temporal Shift Model (TSM) [19] with a ResNet-50 backbone [10], pre-trained on the Multi Moments in Time dataset [22]. This encoder outputs a vector of embeddings with dimension (2048, 8, 7, 7) rather than the original output of (2048, 14, 14). The first dimension of the V2S SAT model embeddings (of size 8) represents the number of frames in each video of our training dataset: Spoken Moments [21]. We pass this encoding to the decoder to output a sequence of units to describe the video.

After making these alterations to the SAT model, we use the pre-trained Tacotron2 model from the baseline task. No alterations are required, since the Tacotron2 model is receiving the same input: a sequence of units.

3.3 ASR Evaluation

In the original work recorded in [12], researchers train a new Automatic Speech Recognition (ASR) model on the synthesized speech, and use that model to transcribe and quantitatively analyze the results from testing data. However, in order to do this, they needed Amazon Mechanical Turk (AMT) workers to manually transcribe the synthesized speech to make a ground truth dataset. They do so because this is the most effective and accurate way to measure the model's ability to synthesize speech. Given the cost and time required for this procedure, however, it can be prohibitive.

In order to mitigate these barriers, we conduct experiments on myriad open-source ASR models. We search for open-source ASR implementations that can perform similarly to a model trained fully on our synthesized speech. We test with 2 open-source ASR models, at 2 different sizes. The first is Wave2Vec 2.0 [5]. In Wav2Vec 2.0, researchers show that a model that first learns representations from speech audio and then is fine-tuned on transcribed speech, outperforms the previous state of the art that requires much more manually transcribed data. We test 2 Wav2Vec models: one that is pre-trained **and** fine-tuned with 960 hours of audio from LibriSpeech [26], and one that is trained with 960 hours of audio from LibriSpeech, and fine-tuned on 10 minutes of audio + ground truth transcriptions from LibriLight [14].

The second set of pre-trained models we test consists of two HuBERT variants [11]. In Hidden-Unit BERT, or HuBERT, researchers aim to improve self-supervised speech representation learning. They do so by utilizing an offline clustering step to provide aligned target labels for a BERT-like prediction loss. They force the model to learn a model over both acoustics and language by applying the prediction loss over the masked regions only. We test the large and extra-large HuBERT pre-trained models.

3.3.1 NLP Scoring Metrics

We employ three metrics to score the language generated by our pipeline: BLEU-4 [27], ROUGE-L [18] and CIDEr [38]. These metrics are all considered standard for evaluating machine synthesis of language.

BLEU-4

The Bilingual Evaluation Understanding (BLEU) [27] score metric evaluates how similar a machine-generated sentence is to a set of ground truth, human-generated sentences. It takes into account the reality that there are often many acceptable ways to translate or caption an input, so it must be able to score a single input sentence against a set of ground truths for each input. The score represents the modified precision of a machine-generated sentence, since precision without any modifications has many pitfalls. Consider, for example, the ground truth caption "The cat is yellow." A machine-generated sentence such as "is is is is is " would technically have a precision of $\frac{5}{5}$, since all five words in the machine-generated sentence appear in the ground truth as well. Such an output, however, should not be rewarded. To combat this problem, the BLEU score adds the following modification to the precision:

Let $count(s, w)$ = the number of times a given word w appears in sentence s from the ground truth corpus of an input. Let $M = max(count(w, s))$ for each word in the output sentence. Then, a given word w is only given credit for M appearances in the output sentence. With this update, an output such as "is is is is is" would be given a BLEU score of $\frac{1}{5}$. The BLEU-4 metric uses this algorithm, but on 4-grams rather than individual words.

ROUGE-L

The Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence (ROUGE-L) [18] finds the longest common subsequence between a corpus of ground truth sentences and the machine-generated sentence. It takes this value

and calculates both the precision p (divide by the number of words in the machine-generated sentence) and recall r (divide by the number of words in the ground truth sentence with the longest common subsequence). Then, the two metrics are combined together to calcuate the ROUGE-L score S:

$$S = \beta * \frac{p * r}{\beta p + r} \tag{3.1}$$

where β represents a tuneable hyperparameter.

CIDEr

The Consensus-based Image Description Evaluation (CIDEr) [38] metric aims to compare a machine-generated sentence against a consensus of the ground truth sentences for a given input. Given a set of ground truth sentences and a candidate machine-generated sentence (all represented as n-grams, where $n = 4$), the CIDEr algorithm takes the following steps:

1. Map every word in the ground truth sentences and candidate sentence to its stem or root form. For example, both "eating" and "eaten" become "eat."

2. Calculate the consensus of the ground truth sentences by calculating a Term Frequency Inverse Document Frequency (TF-IDF) weighting for each n-gram [33]. Intuitively, the TF-IDF weighting encodes the overlap of n-grams in the candidate sentence with n-grams in the reference sentences. It also takes into account the commonness of a given n-gram, and weights an overlap between ground truth and machine-generated for a more common n-gram less heavily than a rare one.

3. Compute the CIDEr_n score with the average cosine similarity between the candidate and ground truth set of sentences.

Chapter 4

Experiments and Results

4.1 Image to Speech

As described in the Methods, we implement the pipeline detailed in [12]. We alter and train the four requisite components, and then compare our quantitative and qualitative results with theirs. Thus, our first experiment investigates the reproducibility of the original results.

4.1.1 Qualitative Results

When qualitatively assessing the outputs of the model, we can see that the ASR model we use for evaluation has a dramatic impact on the perceived performance of the model. When given unit sequences as input, the synthesized speech from the Tacotron2 model tends to add noise and garble some words as it nears the end of a sentence. Usually, this noise is small enough that the output is still intelligible to a human. The ASR models, however, struggle much more with translating these portions of the audio output. This is illustrated in Table 4.1, with additional examples in the Appendix Table A.1. Although these ASR models perform well on human speech, the domain shift into synthesized speech causes additional difficulty. As such, the ASR models manage to capture many of the words from the audio output, but struggles to understand the slightly noisier words from the synthesized speech. This can create a

Human Transcribed	**ASR Transcribed**
the man standing on the side of the road	the man standing on the sidel road
a black, light black, white bird, walking in the water	they dlet light black light firred lalking in the water
a street sign, with people walking on the side of the street	they streak sign with people walking on the side of the street

Table 4.1: Example Human Transcription vs. ASR Transcription for the synthesized audio outputs from the Tacotron2 model. Unit inputs are generated by the I2U SAT model on COCO validation images [20]. The ASR model used here is the HuBERT Large [11].

drastic shift between the ground truth as manually transcribed by a human and the ASR output. For example, a sentence that is not semantically perfect, such as "A black, light black, white bird, walking in the water" becomes completely unintelligible save the last portion with ASR: "They dlet light black light firred lalking in the water." The outputs on the sentence "A street sign, with people walking on the side of the street," on the other hand, are much more similar, with the ASR-transcribed version being "They streak sign with people walking on the side of the street." Outputs such as this highlight why ROUGE-L is a valuable metric for natural language evaluation. Even though "they streak" is very semantically distant from "a street," the rest of the sentence remains true to the context of the image, meaning a longer subsequence in common with the ground truth caption.

Despite the well-formed and human-intelligible nature of most audio outputs from our pipeline, enough outputs have the slightly raspy/garbled quality on which ASR models trained on human speech perform poorly. Even without the context of the accompanying image, audio that is completely misrepresented by this ASR model is understood by a human. It is therefore very feasible that, given an ASR model trained specifically on the output of our model, much more of the overall audio output would be recognized by the ASR model. Full pipeline examples can be seen in Tables 4.2 and A.2.

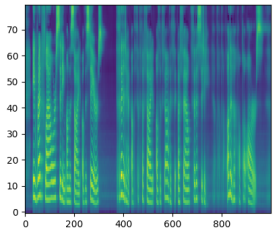
Image	Predicted Units	Waveform & ASR Text
	263, 32, 208, 5, 336, 100, 803, 256, 803, 815, 12, 445, 80, 654, 846, 192, 863, 304, 538, 133...	 a red flower sitting on the side of the stairs

Table 4.2: Example input and output of the entire Image to Speech pipeline from the COCO validation set (input image, unit sequence predicted by the SAT model, output waveform from the Tacotron2 + ASR transcription of audio). A HuBERT Extra Large ASR model is used to generate ASR output. For additional full Image to Speech pipeline examples, see A.2 in the Appendix.

Model	Evaluated on	BLEU-4	ROUGE-L	CIDEr
Hsu et al. [12]	Test	0.186	0.446	0.584
Ours (B = 5)	Validation	0.067	0.300	0.167
Ours (B = 10)	Validation Sample	0.066	0.300	0.163

Table 4.3: Quantitative scoring metrics for the full pipeline, compared to the official baseline as defined by Hsu et al. B = beam size passed to our SAT model when constructing unit sequences. Hsu et al choose a beam size $\in \{3, 5, 8, 10\}$ based on the best SPICE score. Our baseline evaluation results are calculated with validation data from the SpokenCOCO dataset. The beam size = 10 calculation utilized a random sample of 5000 validation images. The evaluation results for Hsu et al are calculated with test data from the SpokenCOCO dataset [12].

4.1.2 Quantitative Results

As can be seen in Table 4.3, the original model from Hsu et al seems to perform markedly better on all scoring metrics tested (namely BLEU-4, ROUGE and CIDEr). However, there is an important caveat to these results: the full model pipeline takes as input a corpus of testing or validation images, gets the units from the SAT Model output, passes the units into the Tacotron2 + Waveglow model, and gets an audio file out. To quantitatively judge via NLP protocols the audio file, the authors trained a full ASR model on the synthesized speech outputs form the Tacotron2 model. They passed the synthesized audio files to the ASR model to get a written version of the caption captured in the audio file. Finally, they used natural language scoring metrics

such as BLEU-4, CIDEr and ROUGE to compare the ground truth written captions from the dataset to the ASR output of the audio file.

With this pipeline, the authors were able to reach their stated values. However, collecting ground truth transcriptions from the synthesized audio is an expensive and time consuming task. Without the hours, workers and resources to do such a task, we turn to open source ASR models to complete the task. For the numbers reported in table 4.3, we used HuBERT Large, a pre-trained ASR model via PyTorch.

With the exception of the ROUGE-L scoring metric, our scores are about one third of the original scores reported in [12]. Their BLEU-4 score, for example, is 0.186, while our is 0.067. Similarly, their CIDEr score is 0.584, while ours is 0.167. The ROUGE-L score, on the other hand, is much more comparable: 0.446 for the original baseline model, and 0.30 for our implementation. ROUGE-L calculates the length of the longest common sequence between a predicted caption and the ground truth caption space. We believe this suits the output from our ASR models, as the text is often accurate and intelligible for a duration, and then gets progressively worse as the noise increases incrementally over the course of the synthesized audio file. A metric like ROUGE-L rewards the long sequence without placing undue negative weight on the devolved, poorly formed portion of the ASR caption. The stark contrast between our qualitative and quantitative results leads us to seek alternate ASR models, or alternate methods entirely, of evaluating out model quantitatively.
This is the motivation for further ASR experimentation and evaluation. Our experiments with other ASR models demonstrate how important the ASR component of this evaluation pipeline truly is: each of these scores, despite being calculated on the same ground truth and predicted values, can change drastically with just a new speech recognition module.

Evaluation Input	BLEU-4	ROUGE-L	CIDEr
Our Unit Sequences	0.282	0.329	0.213
Hsu et al Unit Sequences [12]	0.176/0.274	0.280/0.328	0.121/0.215

Table 4.4: Quantitative scoring metrics from our Image to Unit implementation. Unit sequence scores are calculated by comparing the output of the SAT model to the ground truth unit sequence. Scores are calculated using the MSCOCO validation split for our implementation, and using the MSCOCO test split for theirs. We calculate our scores using a beam size = 5. Their result is Greedy Search/best result from beam size $\in \{3, 5, 10\}$.

Unit-Based Evaluation

Given the difference in performance between our Image to Speech pipeline implementation and the results presented in Hsu et al [12], we compute an additional evaluation procedure. Rather than using the ASR output from our model's synthesized speech as the candidate sentence and the reference written captions as the reference set, we compare the reference and candidate unit sequences. By computing the BLEU-4, ROUGE-L and CIDEr scores of the unit sequences rather than the ASR-transcribed words, we evaluate the SAT model without the potential confounding influence of the Tacotron2 model and our chosen ASR model. This procedure is described and suggested in [12]. The results of our evaluation, as seen in Table 4.4, show that our model matches the baseline set by [12]. Note that the difference in performance likely arises from the use of the validation versus the test split.

These results support the hypothesis that our model's performance compared to the baseline is caused in part by open-source, pre-trained ASR systems that are not equipped to transcribe synthesized speech as effectively as an ASR model trained on these outputs specifically. The unit-based method evaluates only the I2U model's ability to output unit sequences that match the ground truth unit sequences output by the S2U model. These results suggest that our SAT model is performing at the baseline set by Hsu et al.

Human Transcribed	ASR Transcribed
eight skiers skiing and eight fell	ait skers skem and apes fell
the woman has standing and a water and water and s	the woman s standing and a water and lather s
they groub a people walking and walking and eight press and a	they bruple people walking and walking and ait press and it

Table 4.5: Example Human Transcription vs. ASR Transcription for the synthesized audio outputs from the Tacotron2 model given units generated by the **Video to Speech SAT** model. The ASR model used here is the HuBERT Extra Large [11].

4.2 Video to Speech

After altering and re-training the baseline SAT model to process short videos rather than images, we assess the audio output to evaluate the Video to Speech (V2S) pipeline performance. We use the Tacotron2 model pre-trained on units to synthesize speech from the units output by the V2S SAT model.

4.2.1 Qualitative Results

In its current state, the Video to Speech (V2S) model struggles to output intelligible unit sequences. Table A.4 provides examples of the full V2S pipeline. Table 4.5 shows some examples of the ASR transcription output compared to manual human transcription. Here, it is clear that both humans and ASR models struggle to understand exactly what the synthesized audio is saying. However, the model shows it is capable of detecting objects and context from a video. This is evident from the model's ability to generate words and phrases such as "skiers," "water" and "a group of people walking." Future work can improve the V2S pipeline's ability to generate meaningful unit sequences for the downstream Tacotron2 model.

4.3 ASR Evaluation

We evaluate two different types of ASR models: the Wave2Vec and the HuBERT. For each of these models, we assess performance with a smaller and larger fine-tuning corpus. As expected, we find that the larger models (Wav2Vec ASR Base 960 H

Model	BLEU-4	ROUGE-L	CIDEr
Wave2Vec ASR Base 10 M	0.022	0.228	0.028
Wave2Vec ASR Base 960 H	0.057	0.289	0.131
HuBERT ASR Large	0.067	0.300	0.167
HuBERT ASR XLarge	0.071	0.305	0.168

Table 4.6: BLEU-4, ROUGE-L and CIDEr scores for all Automatic Speech Recognition models used to transcribe our Image to Speech model's synthesized speech. The HuBERT ASR X-Large model performs the best, but still does not match the ASR trained model from [12].

and HuBERT ASR Extra Large) perform better than their less fine-tuned counterparts. However, the HuBERT model overall shows significant improvement over either Wav2Vec model. Given this finding, we use the HuBERT models for all ASR use cases in our pipeline.

Future work will expand the suite of ASR models even further to see if another accessible option can perform comparably on the synthesized speech. Future experiments on these ASR models can include fine-tuning the models on synthesized speech rather than additional human speech.

Chapter 5

Conclusion and Discussion

The brunt of this paper focuses on the work done by Hsu et al [12] in their 2020 paper "Text-Free Image-to-Speech Synthesis Using Learned Segmental Units." We build their described baseline model. We evaluate using their reported metrics, and experiment with different Automatic Speech Recognition models in an attempt to replicate their expensive evaluation procedure. We identify HuBERT ASR as a promising open-source model for comprehending and transcribing synthesized audio speech. We also present a novel alteration of the SAT model that allows the model to take a short video as input (in the form of 8 frames) rather than a single image. Although the video version of the SAT model struggles to perform at the same level as the image version, we show that the model is capable of comprehending context, detecting objects, describing the scene, and speaking its learned description. With this work, we introduce a new avenue for speech captioning of videos.

Future work will continue to tune and improve the video SAT model, ideally until its performance is comparable to the image SAT model.

Appendix A

Additional Qualitative Examples

We present additional qualitative examples of our Image to Speech (I2S) and Video to Speech (V2S) pipelines. These examples provide deeper insight into what our model perceives and relays to downstream models at each step of the process.

Human Transcribed	**ASR Transcribed**
a group of sheep standing in a field	eight roof of sheet standing in a field
two white cows standing in the grass	two light pounds standing in the brass
the dog sitting on the bed	the cacks sitting on the bad
a sheet sitting on the desk	a shik sitting on the deft
the light plate with food on the table [inaudible garbled sound]	the light plate with food under table the e la little on of thteblob tebo the vev the te
a dog standing on the ground with a tan	they dod standing underground with eight ten
a man standing next to a horse and a crowd	a man standing next to a horse and a proud
a street sign with stoplights and a street	a street sign with stoplits and a street

Table A.1: Additional example Human Transcription vs. ASR Transcription for the synthesized audio outputs from the Tacotron2 model. Inputs are COCO images. The ASR model used here is the HuBERT Large [11].

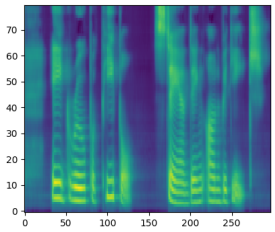
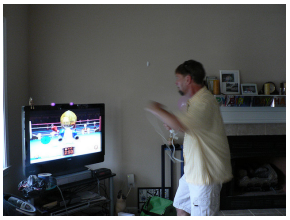
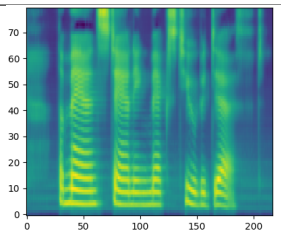
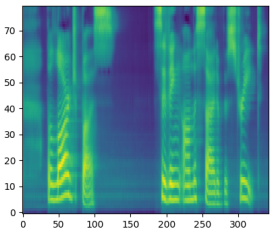
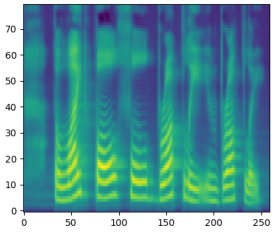
Image	Predicted Units	Waveform & ASR Text
	263, 32, 208, 5, 336, 100, 803, 362, 825, 560, 849, 273, 723, 62, 359, 224, 809, 224, 501, 476...	 a group of people standing on the beach
	263, 32, 208, 263, 336, 501, 791, 728, 971, 728, 141, 536, 113, 580, 827, 1013, 880...	 the man standing next to the desk
	263, 32, 208, 263, 336, 501, 993, 151, 157, 81, 142, 550, 1015, 16, 293, 289, 176, 414...	 the large bus sitting on the side of the street
	263, 32, 208, 263, 336, 501, 993, 151, 81, 749, 133, 715, 825, 560, 476, 224, 414, 284, 44, 341...	 the light ball falled brockley in brodley

Table A.2: Example inputs and outputs of the entire **Image to Speech (I2S)** pipeline from the COCO dataset [20]. A HuBERT Extra Large ASR model is used to generate ASR output.

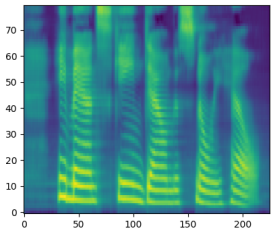
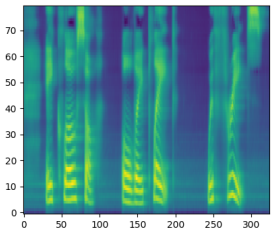
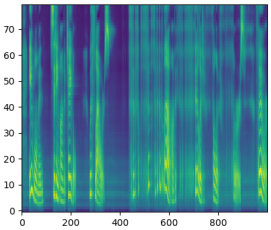
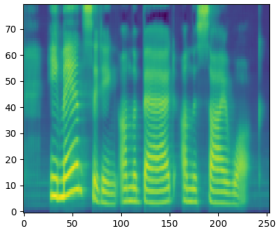
Image	Predicted Units	Waveform & ASR Text
	263, 32, 208, 5, 336, 100, 803, 337, 728, 971, 728, 141, 536, 113, 580, 66, 880, 938, 457, 924, 351, 160...	 he man skeme down the snowly hill he fe he e he ea man tes firess his hame field ffelle h the vol a a a ittle
	263, 32, 208, 5, 336, 100, 803, 256, 560, 873, 986, 96, 651, 538, 135, 522, 135, 639, 606, 360, 794, 892, 872...	 a closeup old a plate with fruts
	263, 32, 208, 5, 336, 100, 803, 256, 803, 81, 831, 156, 719, 329, 89, 119, 936, 48, 336, 5, 208, 5...	 a woman sitting on a table with flowers oe the plou on the fillage full or frs fo
	263, 32, 208, 5, 336, 100, 803, 337, 728, 971, 728, 141, 536, 113, 580, 66, 360, 606, 360...	 he man sitting on the bed on the sidewalk

Table A.3: [Continued] example inputs and outputs of the entire **Image to Speech (I2S)** pipeline from the COCO dataset. A HuBERT Extra Large ASR model is used to generate ASR output.

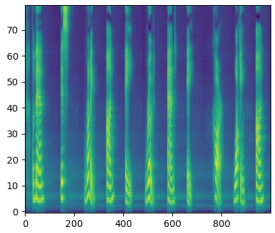
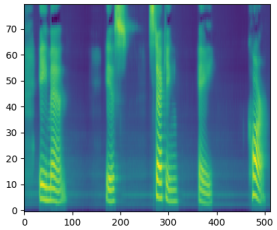
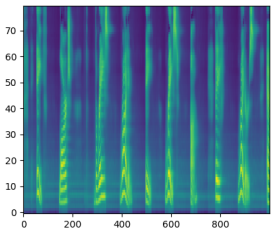
Video Frames	Predicted Units	Waveform & ASR Text
	263, 32, 208, 263, 336, 501, 791, 728, 971, 728, 141, 536, 113, 580, 936, 48, 336, 5, 208...	 the man vis running on a road and a car walking on
	263, 32, 208, 5, 336, 100, 803, 362, 256, 337, 728, 971, 728, 536, 827, 936, 48...	 emen is stacking a car
	263, 32, 208, 5, 336, 100, 803, 362, 256, 560, 846, 476, 5, 208, 32...	 eigt large the race riding ei race com they riders

Table A.4: Example inputs and outputs of the entire **Video to Speech (V2S)** pipeline from the Spoken Moments dataset [21]. A HuBERT Extra Large ASR model is used to generate ASR output.

Bibliography

[1] Imran Alam, Devesh Jalan, Priti Shaw, and Partha Pratim Mohanta. Motion based video skimming. In *2020 IEEE Calcutta Conference (CALCON)*, pages 407–411, 2020.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering, 2017.

[3] Sercan Arik, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. Deep voice 2: Multi-speaker neural text-to-speech, 2017.

[4] Sercan O. Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, Shubho Sengupta, and Mohammad Shoeybi. Deep voice: Real-time neural text-to-speech, 2017.

[5] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

[6] George E. Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):30–42, 2012.

[7] Jeff Donahue, Sander Dieleman, Mikołaj Bińkowski, Erich Elsen, and Karen Simonyan. End-to-end adversarial text-to-speech, 2020.

[8] David Harwath*, Wei-Ning Hsu*, and James Glass. Learning hierarchical discrete linguistic units from visually-grounded speech. In *International Conference on Learning Representations*, 2020.

[9] Mark A. Hasegawa-Johnson, Alan W. Black, Lucas Ondel, Odette Scharenborg, and Francesco Ciannella. Image 2 speech : Automatically generating audio descriptions of images. 2017.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[11] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021.

[12] Wei-Ning Hsu, David Harwath, Christopher Song, and James Glass. Text-free image-to-speech synthesis using learned segmental units, 2020.

[13] Keith Ito and Linda Johnson. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.

[14] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.E. Mazare, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. Libri-light: A benchmark for ASR with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, may 2020.

[15] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions, 2014.

[16] Masatomo Kobayashi, Kentarou Fukuda, Hironobu Takagi, and Chieko Asakawa. Providing synthesized audio description for online videos. In *Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility*, Assets '09, page 249–250, New York, NY, USA, 2009. Association for Computing Machinery.

[17] Kai-Fu Lee. On large-vocabulary speaker-independent continuous speech recognition. *Speech Communication*, 7(4):375–379, 1988. Word Recognition in Large Vocabularies.

[18] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*, 2004.

[19] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding, 2018.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014.

[21] Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. Spoken moments: Learning joint audio-visual representations from video descriptions, 2021.

[22] Mathew Monfort, Bowen Pan, Kandan Ramakrishnan, Alex Andonian, Barry A Mcnamara, Alex Lascelles, Quanfu Fan, Dan Gutfreund, Rogerio Feris, and Aude Oliva. Multi-moments in time: Learning and interpreting models for multi-action video understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[23] Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. Streamlined dense video captioning, 2019.

[24] 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, 2016.

[25] Ian Palmer, Andrew Rouditchenko, Andrei Barbu, Boris Katz, and James Glass. Spoken objectnet: A bias-controlled spoken caption dataset, 2021.

[26] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[27] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[28] Wei Ping, Kainan Peng, and Jitong Chen. Clarinet: Parallel wave generation in end-to-end text-to-speech, 2018.

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

[30] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis, 2018.

[31] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech, 2020.

[32] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech, 2019.

[33] Stephen Robertson. Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation - J DOC*, 60:503–520, 10 2004.

[34] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description, 2016.

[35] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions, 2017.

[36] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A survey on neural speech synthesis, 2021.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[38] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015.

[39] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator, 2014.

[40] Yujia Wang, Wei Liang, Haikun Huang, Yongqi Zhang, Dingzeyu Li, and Lap-Fai Yu. Toward automatic audio description generation for accessible videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[41] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis, 2017.

[42] Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer, 2018.