Synthesizing Spoken Descriptions of Images

Wang, Xinsheng; van der Hout, Justin ; Zhu, Jihua; Hasegawa-Johnson, Mark; Scharenborg, Odette

DOI
10.1109/TASLP.2021.3120644

Publication date
2021

Document Version
Accepted author manuscript

Published in
IEEE - ACM Transactions on Audio, Speech, and Language Processing

Citation (APA)
Wang, X., van der Hout, J., Zhu, J., Hasegawa-Johnson, M., & Scharenborg, O. (2021). Synthesizing Spoken Descriptions of Images. *IEEE - ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3242-3254. Article 9581052. https://doi.org/10.1109/TASLP.2021.3120644

Synthesizing spoken descriptions of images

Xinsheng Wang, Justin van der Hout, Jihua Zhu, *Member, IEEE* Mark Hasegawa-Johnson, *Fellow, IEEE,*
Odette Scharenborg, *Senior Member, IEEE*

***Abstract*—Image captioning technology has great potential in many scenarios. However, current text-based image captioning methods cannot be applied to approximately half of the world's languages due to these languages' lack of a written form. To solve this problem, recently the image-to-speech task was proposed, which generates spoken descriptions of images bypassing any text via an intermediate representation consisting of phonemes (image-to-phoneme). Here, we present a comprehensive study on the image-to-speech task in which, 1) several representative image-to-text generation methods are implemented for the image-to-phoneme task, 2) objective metrics are sought to evaluate the image-to-phoneme task, and 3) an end-to-end image-to-speech model that is able to synthesize spoken descriptions of images bypassing both text and phonemes is proposed. Extensive experiments are conducted on the public benchmark database Flickr8k. Results of our experiments demonstrate that 1) State-of-the-art image-to-text models can perform well on the image-to-phoneme task, and 2) several evaluation metrics, including BLEU3, BLEU4, BLEU5, and ROUGE-L can be used to evaluate image-to-phoneme performance. Finally, 3) end-to-end image-to-speech bypassing text and phonemes is feasible.**

***Index Terms*—Image-to-speech generation, multimodal modelling, speech synthesis, cross-modal captioning.**

I. Introduction

AUTOMATICALLY describing visual scenes using natural language has great potential in many scenarios, e.g., for helping visually-impaired people interact with their surroundings. In recent years, many studies [1], [2], [3], [4], [5], [6], [7], [8] have been conducted in the field of image captioning which aims to automatically generate textual descriptions of images. These neural captioning models follow the encoder-decoder architecture, and are inspired by neural machine translation. The image captioning task has achieved impressive results by integrating various attention mechanisms [4], [5], [9]. However, such text-based image captioning technology cannot be used by people whose language do not have a standard written form. In fact, nearly half of the world's languages does not have a generally-agreed upon written standard [10]. In order to make such image captioning technology accessible to speakers of such unwritten languages, it is necessary to develop technology that automatically creates spoken descriptions of visual scenes, bypassing text. Moreover, such image-to-speech technology [11] has great potential in many other scenarios, e.g., for describing images to visually-impaired people or describing an image when watching a screen is not possible (e.g., when driving a car).

Hasegawa-Johnson et al. [11] first proposed the image-to-speech task that tries to generate spoken descriptions of images without using textual descriptions. In their method, the image-to-speech was decomposed into two stages. The first stage generates speech units, e.g., phonemes, with an image as input. This is also referred to as the image-to-phoneme or image-to-speech unit task. The second stage performs a phoneme-to-speech synthesis process, and completes the image-to-speech task. Importantly, this approach is crucially dependent on the availability of descriptions of the images in terms of sequences of sound units in order to train the first stage. Hasegawa-Johnson et al. compared three different ways of obtaining these sound units: L1 phonemes transcribed by an ASR that was trained on the same language, L2 phonemes transcribed by an ASR that was trained on another language, and so-called pseudo phones that were automatically discovered using an unsupervised acoustic unit discovery system. Only the second two methods would allow for an image-to-speech system that would work for an unwritten language. Unfortunately, only the system based on the L1 phonemes achieved a reasonable performance.

More recently, Hsu et al. [12] used an audio-visual grounding model, named ResDAVEnet-VQ [13], to learn discrete linguistic units from visually-grounded speech. The learned speech units were then used for the image-to-speech task using the image-to-speech unit and speech unit-to-speech approach. Their results show reasonable performance in the image-to-speech task, indicating that the speech units learned by their visually grounded speech method can be used in the image-to-speech task. Effendi et al. [14] also adopted a discrete speech unit discovery model to learn the speech unit representations. Different from [12], in [14], the speech unit discovery model was trained in a self-supervised manner with a speech-only database rather than with a paired image-speech database. Moreover, their speech unit discovery model has an encoder-decoder architecture, in which the decoder takes a speech unit sequence as input and outputs speech representations in the form of spectrograms, allowing this decoder to synthesize speech from the speech units predicted by the image-to-speech

This work was supported in part by the National Key R&D Program of China under Grant 2020AAA0109602 and the Key Research and Development Program of Shaanxi under Grant 2021GY-025. The work of Xinsheng Wang was supported by China Scholarship Council (CSC) (*Corresponding author: Jihua Zhu*)

Xinsheng Wang is with the School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China, and also with the Multimedia Computing Group, Delft University of Technology, 2628 CD Delft, The Netherlands. Email: wangxinsheng@stu.xjtu.edu.cn

Justin van der Hout was with the Multimedia Computing Group, Delft University of Technology, 2628 CD Delft, The Netherlands. Email: justin-vdh@gmail.com

Jihua Zhu is with the School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China. Email: zhujh@xjtu.edu.cn

Mark Hasegawa-Johnson is with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, IL, USA. Email: jhasegaw@illinois.edu

Odette Scharenborg is with the Multimedia Computing Group, Delft University of Technology, 2628 CD Delft, The Netherlands. Email: O.E.Scharenborg@tudelft.nl

unit model. Both these approaches outperformed the pseudo-phone-based method in [11].

So several efforts [11], [12], [14] have pursued the goal of image-to-speech synthesis using the consecutive steps of image-to-speech unit and speech unit-to-speech, and with reasonable results. However, many questions still remain. This paper focuses on three. First, although image-to-speech is a new task, it shares the same idea as text-based image captioning methods, both of which follow the basic structure of neural machine translation. There are, however, many more models proposed for image captioning than for image-to-speech; none of which have been investigated for the image-to-speech (unit) task. So the first question we aim to answer in this work is: In how far can current image captioning methods be used for the image-to-speech task, and more specifically the image-to-phoneme task? To that end, we implemented several representative image captioning models in the image-to-phoneme system. The image-to-phoneme model proposed in [11] was re-implemented to serve as the baseline to which the image-to-text-based systems were compared.

In the original image-to-speech paper [11], the evaluation of the system was only carried out for the image-to-phoneme task. BLEU score and speech unit (phoneme) error rates were adopted as the evaluation metrics. Our recent work [15] investigated the suitability of several metrics for the objective evaluation of the image-to-phoneme task by correlating these metrics with human ratings. BLEU4 was found to be the best metric. However, in this work, only one model, i.e., a re-implementation of the model developed by [11] was considered. Therefore, the second aim of the current paper is to further investigate objective measures with different models. Here, we extend our previous work by investigating the suitability of these objective metrics on the evaluation of the image-to-phoneme task by correlating the scores for the different image captioning models on the image-to-phoneme task with the scores on these metrics after conversion of the automatically generated phoneme sequences to words.

Third, as explained above, the performance of the speech unit-based method crucially depends on the quality of the used speech units to train the image-to-speech unit stage. So the third question we aim to answer here is whether the image-to-speech task can be realized by an end-to-end model that bypasses the need for both text and intermediate speech units. To that end, we propose an end-to-end image-to-speech model which can generate spoken descriptions directly from images without using any intermediate speech units.

In this work, we focus on both the image-to-phoneme task and the image-to-speech task. The contributions of this work are as follows:

- Experiments on various image captioning models that were implemented for the image-to-phoneme task showed that image captioning methods can have a good performance on the image-to-phoneme task.
- Analysis of various evaluation metrics' effectiveness on the image-to-phoneme task showed that BLEU3, BLEU4, BLEU5, and ROUGE-L evaluations of a phoneme string correlate well with objective evaluations of the resulting text output.
- For the first time, an end-to-end image-to-speech method was proposed, which demonstrated that generating spoken descriptions for images while bypassing text and intermediate speech units is feasible.

The rest of the paper is organized as follows: Section II reviews related works on image captioning and visual-speech multi-modal learning. Section III introduces several image captioning models that will be re-implemented in the image-to-phoneme task. Section IV describes the proposed end-to-end image-to-speech method. Section V and VI present the results of the image-to-phoneme task and the end-to-end image-to-speech method, respectively. Section VII discusses the results of the experiments and proposes research directions for the future. Finally, Section VIII concludes this paper.

II. Related works

A. Image captioning

Earlier image captioning approaches were retrieval-based or template-based methods. In retrieval-based methods, the caption of an image is obtained by retrieving one or a set of sentences from a pool of existing sentences [16], [17]. The template-based method is normally based on the outputs of an object detector or attribute predictor to compose a sentence by adding these objects or attribute words to a caption template [18], [19]. Although both these methods usually lead to grammatically correct and fluent captions, the obvious disadvantage is that the number of different captions that can be generated is limited due to the use of pre-existing sentences and templates.

In recent years, inspired by the development of neural machine translation, the neural-based encoder-decoder paradigm has been the basic framework used for image captioning. In [20], [21], the image is encoded by a Convolutional Neural Network (CNN), and a Recurrent Neural Network (RNN) is used as a sentence decoder to generate a text description of the input image. This encoder-decoder framework has shown to achieve promising results [20]. However, as is well known, an image contains rich information, much of which typically is not described by humans when captioning the image (e.g., the clouds in the sky or trees in the background). Motivated by the visual attention mechanism of primates and humans, the attention mechanism has been successfully used in automatic image captioning systems, which led to large improvements [1], [4], [22], [23], [24].

In [1], the authors integrate the attention mechanism into the decoder to encourage more interactions between the image and generated sentences by selecting the relevant image regions during the decoding process. After that, many efforts [4], [22], [25] have been made to boost the image captioning performance by designing more effective architectures of the attention mechanisms. For instance, in [22], an adaptive attention mechanism is introduced to decide when to activate the visual attention. The Attention on Attention model proposed in [4] is designed to determine the relevance between the attention result and the query.

Most recently, due to its success in natural language processing (NLP) [26], the Transformer model has been implemented in the image captioning task [3], [4], [5], [27]. In [4], a Transformer-like encoder was paired with an LSTM decoder. $\mathcal{M}^2$ Transformer [5] is completely based on the Transformer. Compared to the vanilla Transformer architecture, in the $\mathcal{M}^2$ Transformer, the encoding and decoding layers are connected in a mesh-like structure to exploit multi-level relationships among image regions. In [9], the authors integrate the proposed X-Linear attention block with the Transformer architecture for the image captioning task.

In this paper, several representative models of image captioning [1], [4], [20], [25], [9], which vary from a classical method to a state-of-the-art method, are adopted in the image-to-phoneme task. Details of these adopted models can be found in Section III-B.

B. Cross-modal learning between visual and speech

Inspired by human infants' ability to learn spoken language by listening and paying attention to the concurrent speech and visual scenes, recently a new research area emerged in which speech representations are learned grounded by corresponding images [28], [29], [30], [31], [32], [33], [34], the so-called visual-grounded speech learning task. For instance, in [29], images and the corresponding spoken captions were mapped into a common embedding space by an image encoder and a speech encoder respectively. In the embedding space, the image representation can work as the supervision information to train the speech encoder. This task which relies on a matching relationship between images and their corresponding spoken descriptions spawned several other cross-modal tasks between visual and speech, i.e., the segmentation of the objects in an image and keywords in an utterance [28], [35] and multimodal word discovery [36], [37]. Most recently, Wang et al. [38], [39] proposed the S2IGAN model to generate images based on spoken descriptions. In this speech-to-image generation model, the speech representations are also learned with the grounding of corresponding images.

III. Image-to-phoneme

To investigate whether image captioning models can be used in the image-to-phoneme task, several representative image captioning models were implemented for the image-to-phoneme task. Additionally, we used our re-implementation [15] of the original image-to-phoneme system proposed in [11] as the baseline system. Details of those image captioning models and the re-implemented image-to-phoneme system will be introduced in this section.

A. Re-implementation of the image-to-phoneme model

Our baseline image-to-phoneme model [15] is based on the extensible Neural Machine Translation Toolkit (XNMT) [40]. The image-to-phoneme model is an attention-guided encoder-decoder architecture. The encoder takes image features as input, and the decoder outputs the predicted phoneme sequence. The encoder uses 3 layers pyramidal LSTM with 128 units. The attender uses a multi-layer perceptron with a state dimension of 512 and a hidden dimension of 128. The decoder is a 3 layer LSTM with 512 units followed by a multi-layer perceptron with a hidden dimension of 1024 working as a transformation between the outputs of LSTM and a final softmax layer. Compared to the original image-to-phoneme model [11], the number of encoder layers was increased from 1 to 3 and the attender state dimension was increased from 128 to 512, which led to a performance increase. More details of this re-implementation and a comparison of its performance with the original model of [11] can be found in [15]. For convenience, this re-implemented image-to-phoneme model is referred to as R-I2P hereafter.

B. Image captioning methods

Several image captioning models [1], [4], [20], [25], [9], including a basic encoder-decoder architecture (Neural Image Caption [20]), a standard attention-guided model (Show, Attend and Tell [1]), and several state-of-the-art models (Updown model [25], Attention on Attention Model [4], and X-Linear Attention Network [9]), were implemented for the image-to-phoneme task. All these models use an encoder-decoder architecture but with different attention mechanisms or none:

Neural Image Caption (NIC) [20] is a basic image captioning model, which uses a deep CNN to encode the input image and uses an LSTM to decode the caption sequence. In this system, the image is only shown to the LSTM at the beginning and no attention mechanism is adopted.

Show, Attend and Tell (SAT) [1] uses the soft attention mechanism proposed in [1], which is a standard attention mechanism. In this model, the spatial attention mechanism on the image feature map is used to automatically focus on salient objects when inferring the next word to be generated.

Updown model [25] combines bottom-up and top-down attention that enables attention to be calculated at the level of objects and other salient image regions. The "bottom-up" refers to the purely visual feed-forward attention mechanisms and the "top-down" refers to attention mechanisms driven by non-visual or task-specific context. In this model, the bottom-up mechanism is based on the Faster-RCNN [41] to detect the interesting regions of the images, and the top-down mechanism determines the feature weights of different image regions.

Attention on Attention model (AoANet) [4] uses a multi-head attention mechanism that is similar to the attention mechanism in the Transformer model to encode the image features. Different from the original Transformer encoder, in AoANet, the feed-forward layer is replaced by the proposed Attention on Attention module (AoA). In the decoder, the AoA module is incorporated with the LSTM to predict word sequences. The AoA module is designed to determine the relevance between the attention results and the query.

X-Linear Attention Network (X-LAN) [9] utilizes a new attention block, referred to as the X-Linear attention block, which adopts bilinear pooling to capture the 2nd order interaction between the input single-modal or multi-modal features. In the X-Linear attention block, both the spatial and channel-wise bilinear attention distributions are considered. A variant of X-LAN, named X-Transformer, is obtained by plugging X-Linear attention blocks into the Transformer.

C. Training and inferring methods

All the models that were originally designed for image captioning and are adopted in this paper were trained with a phoneme-level cross-entropy loss. Both reinforcement learning and beam search strategies have shown good performance on the image captioning task [6], [42], therefore both strategies were also investigated in the image-to-phoneme experiments. Image captioning models are usually trained using the cross-entropy loss, while they are evaluated using discrete and non-differentiable NLP metrics such as BLEU, ROUGE, METEOR, or CIDEr. Therefore, a discrepancy could occur between the training objective function and the evaluation metrics. Reinforcement learning that directly optimizes on the metrics showed a good performance for image captioning [6]. Here, we adopt the self-critical sequence training (SCST) method proposed in [6]. Due to the best correlation with human rating [15], BLEU4 was adopted as the reward. During the inference process, the auto-regressive model normally greedily selects the most probable output of the next step. Here, we investigate the beam search method which maintains a list of the N most probable sub-sequences generated so far, generates posterior probabilities for the next word of each of these sub-sequences, and then again prunes down to the N-best sub-sequences. The beam search method was found to provide a boost in the performance of image captioning [6], [42].

D. Evaluation metrics

Ideally, the image-to-speech task should be evaluated in terms of the generated speech signal. However, the generated speech from the phoneme sequence will be affected by the particular phoneme-to-speech system that is used. Because we want to evaluate the content of the caption irrespective of the naturalness or intelligibility of the generated speech, we evaluate the image-to-phoneme task on the level of the generated phoneme sequences.

The evaluation metrics we use are several popular metrics for the image captioning tasks, i.e., BLEU [43] (bilingual evaluation understudy), METEOR [44] (Metric for Evaluation of Translation with Explicit ORdering), ROUGE-L [45] (Recall-Oriented Understudy for Gisting Evaluation with the Longest Common Subsequence), and CIDEr [46] (Consensus-based Image Description Evaluation). In calculating the metrics, we follow the standard approaches as used in image captioning evaluation, in which all captions of each image are used as reference sentences.

E. Dataset

One of the key reasons behind the image-to-speech task is to allow speakers of unwritten languages benefit from image captioning systems. However, unwritten languages are not only low-resourced, but also under-researched. Currently, there are no appropriate databases of unwritten languages that can be used in the image-to-speech task. In this paper, the well-resourced language, i.e., English, is adopted as the working language, and treated as if it were a low-resource, unwritten language. The benefit of using English as the working language is that it allows for easy evaluation of the captions. Specifically, the Flickr8k [16] and its associated Flickr-Audio corpus [47] are used in this research. The Flickr8k image database contains 8,000 images from Flickr, and each image has five textual captions, which have been obtained using Amazon Mechanical Turk (AMT) [16]. The audio corpus, which was also collected via AMT [47], consists of speech recordings of the textual captions. The utterances were forced aligned with their corresponding phonemic transcriptions (in ARPABET) using the Janus Recognition Toolkit [48].

We use the standard way to split the Flickr8k: 6,000 images for training and 1,000 images both for development and test set. However, a few sentences could not be forced aligned, therefore, we eventually used 5,662 images, 961 images, and 952 images for the training set, validation set, and test set, respectively. Each image has up to 5 phoneme sequences, and the final training set, validation set, and test set have 28,205, 4,741, and 4,741 phoneme sequence captions, respectively.

In the original I2P model in [11], image CNN features were obtained using VGG-16 [49] pre-trained on ImageNet [50] by scanning the penultimate convolutional layer of VGG-16 in raster-order, resulting in 196 sequential feature vectors of dimension 512 of each image. We followed this approach in our baseline model. The other image captioning models, to compare them fairly, are based on the feature extracting method proposed in [25], in which the Faster-RCNN [41] model pre-trained on ImageNet [50] and Visual Genome [51] was adopted.

IV. End-to-end image-to-speech

The proposed end-to-end model, referred to as the Show and Speak (SAS) model, is based on an encoder-decoder framework. The encoder plays the same role as in the image captioning systems and the image-to-phoneme systems, and encodes an input image to an embedding space, where the image is represented by a sequence of feature vectors. Then, the decoder takes these image feature vectors as input to synthesize a spoken description of the image. In the proposed framework, the speech is represented by a spectrogram. The architecture of the proposed method is shown in Fig. 1 and will be explained in detail below.

A. Encoder

The structure of the encoder is shown in the left-most column in Fig. 1. Given an image, the encoder obtains a sequence of image feature vectors $\{v_1, v_2, ..., v_l\}$ of l object regions from the image using a pre-trained object detector. Here, following [25], the Faster-RCNN [41] model pre-trained on ImageNet [50] and Visual Genome [51] is adopted to extract image features of $l = 36$ object regions. The extracted feature vectors of one image are presented as $\{f_1, f_2, ..., f_l\} \in \mathbb{R}^{l \times d}$, where $d = 2048$ is the feature dimension. For each local feature f_i, the pre-trained Faster-RCNN [25] provides its position in the image, predicts the class label, and computes its confidence score (possibility), which are represented as p_i, c_i, and s_i respectively. Specifically, $p_i \in \mathbb{R}^5$ consists of four bounding

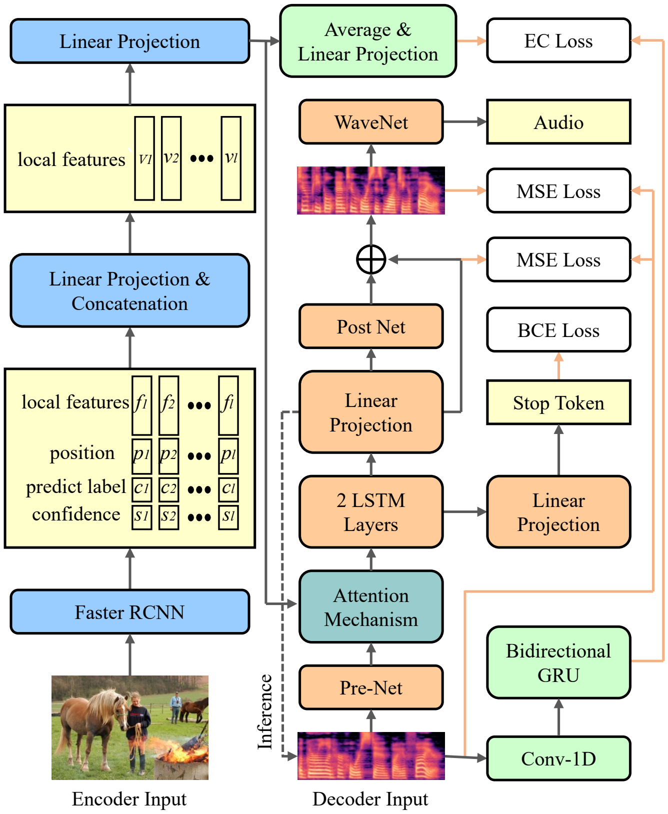


Fig. 1: Architecture of the Show and Speak (SAS) model.

box coordinate values, i.e., top left (x, y) and bottom right corner (x, y), and one ratio value of the bounding box area to the image area. The predicted class label $c_i \in \mathbb{R}^{1601}$ is a one-hot vector, and its corresponding confidence score s_i is a real value. One advantage of adding this information is to correct the errors caused by the detection model. An ideal detected region should contain only one object and should contain it completely, in which case the confidence score would be highest. In contrast, in a badly detected case, e.g., several objects appear in the same detected region. This will result in a low confidence score, and the SAS decoder can learn to use this low score to suppress the role of the feature from this region. Following [52], the image feature v_i is obtained via

$$v_i = f_i \oplus \left[FC\left(p_i \oplus c_i \oplus s_i\right)\right], \tag{1}$$

where $\oplus$ means concatenation and FC is a linear projection with 1024 units. Then the image is represented as $V = \{v_1, v_2, ..., v_l\} \in \mathbb{R}^{36\times3072}$. Finally, in order to create image representations that are more consistent with spoken captions, the image features are passed through two linear transformation layers of 1025 and 512 units respectively to get image embeddings with the dimension of 512. The decoder is trained (parameters of the pre-trained Faster-RCNN are fixed) in the encoder-decoder system with the extra embedding constraint that will be introduced in Section IV-C.

B. Decoder

The structure of the decoder is shown in the middle column of Fig. 1 (from the decoder input to the spectrogram before the WaveNet). The decoder takes the image feature sequence output from the encoder as input to synthesize speech spectrograms in an autoregressive way. The speech is represented by 80 channel log mel spectrogram computed through a short-time Fourier transform (STFT) with 50 ms frame size and a 12.5 ms frame hop. The decoder architecture follows the structure of Tacotron2's decoder [53]. Specifically, the generated spectrogram frame from the previous time step passes through a Pre-Net and is then concatenated with an attention context vector before being passed through two LSTM layers. The attention context vector is obtained from the encoder output with the location-sensitive attention [54], and the Pre-Net consists of 2 fully connected layers both of which have 256 hidden units. The output of the LSTM is concatenated with the attention context vector and then passed through a linear projection to generate the spectrogram frame of the next time step. Then, the generated spectrogram passes through a Post-Net, which consists of 5 convolutional layers with 512 filters, to get a spectrogram residual that is added to the spectrogram before the Post-Net in an element-wise way, achieving the final generated spectrogram. Finally, the generated spectrograms are inverted into time-domain waveform samples via a modified version of WaveNet [55] in [53].

C. Objective function

Following the objective function in Tacotron2 [53], mean squared error (MSE) is used to optimize the generation of spectrograms before and after Post-Net. We denote the synthesized spectrograms before and after Post-Net by X^b and X^a respectively, and denote the ground-truth spectrogram by X, the loss function for optimizing the spectrogram is defined as

$$\mathcal{L}_s = \frac{1}{n}\sum_{i=1}^{n}\left(\left\|X_i^a - X_i\right\|^2 + \left\|X_i^b - X_i\right\|^2\right), \tag{2}$$

where n is the batch size.

To allow for the model to dynamically determine the length of the predicted spectrogram instead of synthesizing a fixed-length sequence, a "Stop Token" prediction module that is similar to the module in [53] is adopted in the proposed framework. Specifically, the concatenation of the decoder LSTM output and attention context vector passes through a linear transformation layer to obtain a scalar followed by a sigmoid activation, resulting in a probability which predicts whether the output sequence has completed or not. The corresponding loss function is binary cross-entropy (BCE) loss, which is defined as

$$\mathcal{L}_{st} = \frac{1}{m \cdot n}\sum_{j=1}^{n}\sum_{i=1}^{m}\left[y_{ij} \cdot \log z_{ij} + \left(1 - y_{ij}\right)\log\left(1 - z_{ij}\right)\right], \tag{3}$$

where y is the label to indicate whether the frame is a stop token or not. z is the output of the sigmoid activation layer. m and n indicate the length of the spectrogram sequence and the batch size respectively.

In parallel to the prediction of the spectrograms and stop tokens, an image embedding constraint (EC) loss is introduced to penalize any component in the image embedding that cannot be predicted from the spoken caption, i.e., any component of the image embedding that is semantically independent of the caption. The rounded boxes with the green background in Fig. 1 show the operations for the image embedding constraint. The image global feature vector, u, is obtained by averaging the encoder outputs, and a linear transformation layer is implemented on the averaged vector to get the final image global feature vector that is used to calculate the EC loss. The neural network structure to get the speech embedding vector is similar to the speech encoder in [38]. Specifically, the ground-truth speech spectrogram first passes through a 1-D convolutional layer, and the fixed-length speech feature vector, r, is obtained by averaging the output of a two-layer bi-directional gated recurrent units (GRU). The matched image-speech vectors should be close to each other, while at the same time different from other unmatched vectors. To that end, we use the Masked Margin Softmax (MMS) method [31] to obtain the EC loss. The EC loss is defined as

$$\mathcal{L}_{ec} = -\frac{1}{n}\left(\sum_{i=1}^{n} \log \frac{e^{S_{ii}}}{\sum_{j=1}^{n} e^{S_{ij}}} + \sum_{j=1}^{n} \log \frac{e^{S_{jj}}}{\sum_{i=1}^{n} e^{S_{ij}}}\right) \quad (4)$$

where n is the batch size, S is a similarity matrix between each feature vector pair within a batch. Each element in the similarity is defined as the dot product between the feature vectors:

$$S_{i,j} = u_i \cdot r_j. \quad (5)$$

The total loss for training the SAS model in an end-to-end way is given by

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_{st} + \lambda \mathcal{L}_{ec}, \quad (6)$$

where λ is a hyperparameter to balance the image embedding constraint. The value of λ is experimentally set as 0.25 out of $\{0.1, 0.25, 0.5, 0.75, 1.0\}$.

D. Evaluation metrics

The image-to-speech task is evaluated in terms of how well the synthesized spoken caption describes its corresponding image. However, it is difficult to directly evaluate the spoken captions. As explained in Section III-D, we want to evaluate the content of the generated speech rather than the speech signal. Therefore, in order to objectively evaluate the image-to-speech task, the synthesized speech is automatically transcribed to text. To that end, an automatic speech recognition (ASR) system[1] built with Kaldi [56] is used. The ASR system consists of a hybrid factorized time-delay neural network (TDNN-F) [57] acoustic model (AM) and a four-gram language model (LM), both trained using the 960-hour Librispeech English database [58].

The transcribed textual captions are then evaluated using the evaluation metrics for image captioning [4], [5]: BLEU4, METEOR, ROUGE, and CIDEr. Because the evaluation is performed on the textual captions that are transcribed from the spoken captions using the ASR system, higher scores of those metrics will also to a certain extent reflect a better quality of the synthesized speech as a worse quality of the synthesized speech would seriously affect the accuracy of the ASR system.

E. Training Details

We train the SAS network using the Adam optimizer with a warmup in the first 4,000 iterations, and a learning rate that decreases with a continuous exponential decrease from 2e-3. The standard neural sequence-to-sequence training procedure, referred to as the teacher-forcing method, feeds the decoder with the ground-truth spectrogram. In the inference stage, this training method could yield errors that can accumulate quickly along the generated sequence due to the discrepancy between training and inference. Here, we adopt the scheduled sampling [59] to alleviate this problem.

F. Dataset

Following the previous experiments on the image-to-phoneme task, Flickr8k [16] is also used in this end-to-end experiment, and we use the same training, validation, and test set splits as in the previous experiments. The speech recordings of this database come from 183 different speakers, making speech synthesis a challenging task. Here, to eliminate the impact of speakers, we adopt a text-to-speech (TTS) system [53] trained on a single speaker to synthesize the spoken captions. This TTS system is pre-trained on LJSpeech [60] which consists of 13,100 audio clips recorded from a single speaker. While the multi-speaker speech synthesis is not a main concern in the current work, we still conduct a further experiment with the original recorded multi-speaker spoken descriptions to investigate how well the proposed model can perform on the multi-speaker natural speech dataset (see Section III-E).

V. Results on the Image-to-phoneme Task

In this section, we first investigate which evaluation metrics can evaluate the image-to-phoneme task well. Subsequently, the different image-to-phoneme models are compared on the image-to-phoneme task in terms of those metrics that were found most suitable for evaluating the image-to-phoneme task (Section V-B). Finally, we present the results of the experiments on the effect of beam search and reinforcement learning on the image-to-phoneme task.

A. Evaluation of metrics

In the previous human rating experiments [15], it has been demonstrated that BLEU3, BLEU4, BLEU5, and ROUGE-L show a higher correlation with the human ratings than other metrics. However, these results are obtained on only one model. To further investigate the usefulness of these metrics for the speech-to-phoneme task, we 1) calculate the scores of the different metrics on the generated phoneme sequences from several models. However, instead of correlating these

[1] https://kaldi-asr.org/models/m13

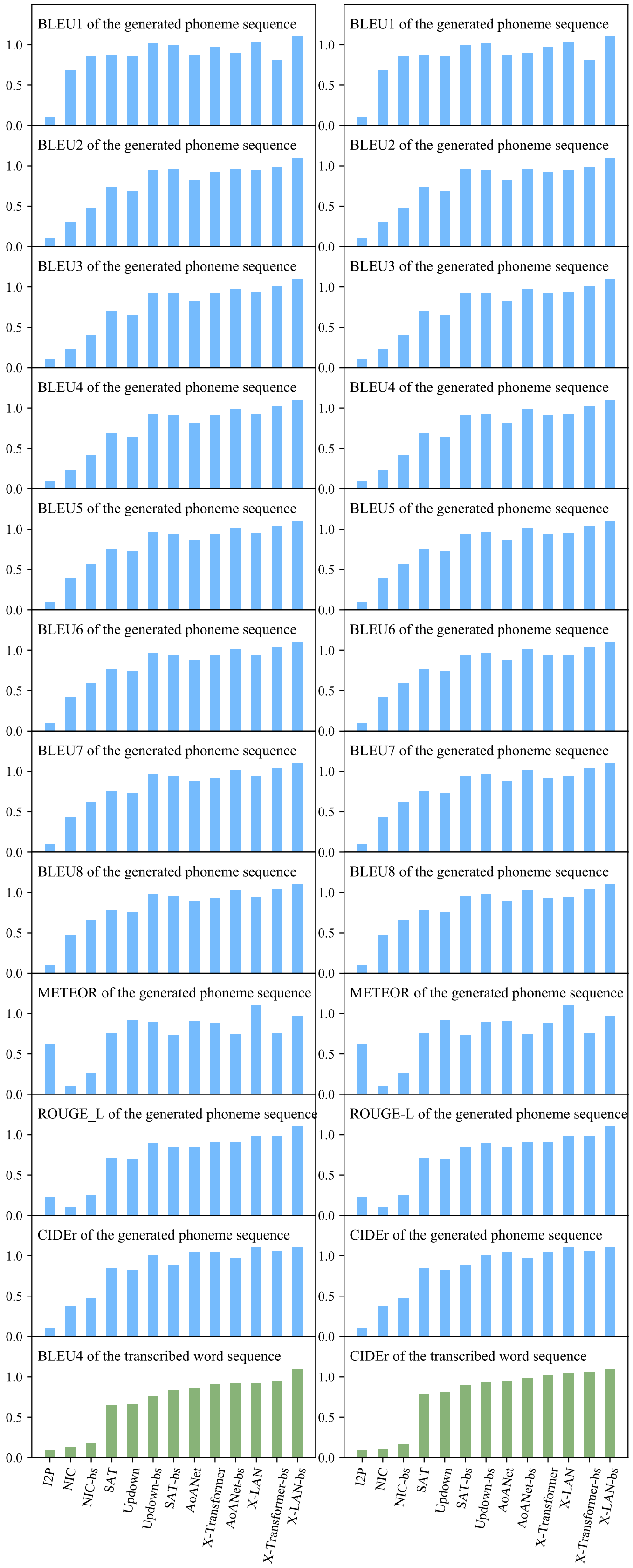


Fig. 2: Comparison of the image-to-phoneme scores and the image-to-text scores in terms of the different evaluation metrics. On the left, the model lists are ranked in an ascending order based on the BLEU4 scores on the image-to-text results. On the right, the model lists are ranked in an ascending order based on the CIDEr scores on the image-to-text results.

with human ratings, which is a time-consuming and costly enterprise, we take advantage of existing evidence of the correlation between the image captioning metrics on text and human ratings, and we 2) compare them with the scores of the same metrics computed on word sequences that are automatically derived from the generated phoneme sequences in the next subsection.

We first compute the scores of all the metrics, i.e.,BLEU1-BLEU8, METEOR, ROUGE-L, and CIDEr, on the generated phoneme sequences of the different image-to-phoneme models (see Section III), including the baseline R-I2P model. Here, image-to-phoneme models are evaluated with and without beam search respectively. For instance, “NIC-bs” means the NIC model was implemented with beam search, while no beam search exists in the "NIC". We then use the wFST from [15] to convert all generated phoneme sequences of all image-to-phoneme models to text. We refer to this output as image-to-text. Please note that the used WFST is rather strict as it does not allow for phoneme insertions, substitutions or deletions. Consequently, isolated phonemes may occur in the final word sequence. In [15], these isolated phonemes were removed from the word sequence for the convenience of the human rating experiments. Here, however, we treat these phonemes as errors of the model, and consequently, their existence reduces the measured quality of the text caption produced by any given model.

In order to investigate which metrics, when applied to the phoneme string, correlate well with objective evaluations of the corresponding word string, we first evaluate the performance on image-to-text of each model. We take two metrics that are able to evaluate text captions well, i.e., BLEU4 and CIDEr, of which BLEU4 is the most commonly used metric out of BLEU@N in image captioning [6], [61] and CIDEr is a metric that well correlates with human judgment [46]. We use these two metrics to compute the scores for all image-to-text output. So for every model, we now have a score for each of the metrics on the image-to-phoneme output and a score for BLEU4 and CIDEr on the image-to-text output. For each metric, a higher score means a better performance of the model.

The different scores calculated on the transcribed text results can be ranked in order of increasing performance, i.e., a ranking of the models from worst-to-best performing model. A good metric for the image-to-phoneme task, then, should be able to show a similar worst-to-best ranking as BLEU4 and CIDEr on the image-to-text output. Therefore, in order to investigate which of the metrics is able to evaluate the image-to-phoneme task, we plot the scores of the metrics on the image-to-phoneme task, separately for each metric, and ranked in order of increasing score of the BLEU4 for on the image-to-text output (see the left part of Fig. 2; green, bottom plot) and ranked in order of increasing performance on the CIDEr metric for the image-to-text output (see the right part of Fig. 2; green, bottom plot).

For the convenience of display, the scores in each histogram are normalized to $[0.1, 1.1]$. Comparing the rankings of the two text-based metrics, we see they are similar and only have two pairs of adjacent models, i.e., Updown-bs/SAT-bs

TABLE I: Performance of the different models on 1) the image-to-phoneme task, where the generated phoneme strings are evaluated using BLEU3 (B3), BLEU4 (B4), BLEU5 (B5), and R(OUGE-L); 2) the image-to-text output where the text results are evaluated using BLEU4 (B4), M(ETEOR), R(OUGE-L), and C(IDEr), respectively. Bold indicates the best result of each metric. Higher scores are better. Effendi et al. only reported the BLEU4 score.

	Phoneme results				Transcribed text results			
Methods	B3	B4	B5	R	B4	M	R	C
R-I2P	46.4	36.1	24.6	49.3	8.1	15.7	30.3	21.7
Effendi et al. [14]	—	46.2	—	—	—	—	—	—
NIC [20]	48.4	38.3	31.1	48.3	8.4	15.8	32.7	22.0
SAT [1]	55.4	46.3	39.1	53.4	12.8	19.4	37.9	40.3
Updown [25]	54.8	45.6	38.4	53.2	12.9	19.8	37.7	40.7
AoANet [4]	57.2	48.6	41.6	54.5	14.7	20.6	39.0	44.4
X-LAN [9]	**58.9**	**50.4**	**43.4**	**55.6**	**15.2**	**21.2**	**39.7**	**46.9**
X-Transformer [9]	58.7	50.1	43.1	55.1	15.0	20.9	39.5	46.2

and X-Transformer/AoANet-bs, with different orders. However, because the scores for Updown-bs/SAT-bs and for X-Transformer/AoANet-bs are very similar for both metrics, this reversal of the exact ranking is not important.

Comparing the different metrics on the image-to-phoneme output (blue panels) with the ranking of BLEU4 and CIDEr on the image-to-text output, we can see that the different metrics do not show the exact same ranking in performance of the models as the bottom panels. As shown in Fig. 2, the ranking of the performance of the models in terms of the METEOR metric shows large differences from that of BLEU4 and CIDEr on the image-to-text output in the bottom panels. In the BLEU@N scores, when the N is small, i.e., less than 3, the differences are also obvious. Those results indicate that METEOR, BLEU1, BLEU2, and BLEU3 are not suitable metrics to evaluate the image-to-phoneme output.

From Fig. 2, we can see that BLEU3, BLEU4, BLEU5, BLEU6, ROUGE-L and CIDEr produce a similar increasing trend to the text-based scores, indicating they are better metrics to evaluate the image-to-phoneme task.

So, in line with the human rating experiment in [15], we also observe here that BLEU3, BLEU4, BLEU5, and ROUGE-L seem to be the best metrics to evaluate the image-to-phoneme task. Henceforth, these objective measures will be used to evaluate and compare the different models on the image-to-phoneme task.

B. Comparison of the performance of the different models on the image-to-phoneme task and image-to-text output

The performance of the different models is compared on both the image-to-phoneme task and the image-to-text output. Please note that the commonly used metrics for the image-to-text task are BLEU4, METEOR, ROUGE, and CIDEr (see Section IV-D). Therefore, we will evaluate our image-to-text output (i.e., the word sequences converted from the generated phoneme sequences) with these four metrics. The results are presented in Table I. The left side of the table presents the evaluation of the generated phoneme sequences of the image-to-phoneme models. The right side of the table presents the evaluation of the word sequences converted from the generated phoneme sequences (the image-to-text output). In addition to the baseline R-I2P model, we also add the Image2Text model proposed by Effendi et al. [14] to this table, as they also performed their Image2Text model on the same dataset, which allows us to compare directly. As can be seen, all models that were originally designed for the image captioning task and the model proposed by Effendi et al. achieve a better performance than the baseline R-I2P model in [14] on both the generated phoneme sequences and the converted text output.

TABLE II: Effect of reinforcement learning and beam search. + bs means the beam search was adopted during inference, + rf means the reinforcement learning was used to fine-tune the model during training. Bold indicates the best result of each metric, and italics indicate worse results compared to the corresponding model without reinforcement learning and beam search.

	Phoneme results				Transcribed text results			
Methods	B3	B4	B5	R	B4	M	R	C
NIC + bs [20]	51.0	41.6	34.8	49.5	8.8	16.1	32.7	23.4
SAT + bs [1]	58.7	50.1	43.2	54.5	14.5	19.7	38.2	43.0
Updown + bs [25]	58.9	50.5	43.7	54.9	13.8	20.1	38.4	44.1
AoANet + bs [4]	59.6	51.4	44.8	55.1	15.1	*20.4*	*38.7*	45.3
X-LAN + bs [9]	61.5	53.4	46.8	56.6	16.7	**21.4**	40.2	48.4
X-Transformer + bs [9]	60.1	52.0	45.4	55.6	15.4	*20.6*	39.2	47.5
NIC + rf [20]	55.6	45.5	37.8	50.4	*8.0*	*15.7*	*32.1*	23.3
SAT + rf [1]	59.4	50.1	42.6	53.4	*11.2*	*17.5*	*35.2*	*33.1*
Updown + rf [25]	61.5	52.5	44.9	53.2	*11.9*	*18.2*	*36.5*	*35.7*
AoANet + rf [4]	64.5	56.4	49.6	56.2	16.1	*20.3*	39.8	47.0
X-LAN + rf [9]	64.4	56.7	50.1	**57.8**	18.3	**21.4**	40.8	51.3
X-Transformer + rf [9]	63.4	55.3	48.6	56.9	17.2	*20.8*	40.4	48.7
NIC + rf + bs [20]	55.6	45.6	37.8	50.4	*8.0*	*15.5*	*32.1*	22.8
SAT + rf + bs [1]	59.7	50.5	43.1	53.3	*11.2*	*17.5*	*35.1*	*32.7*
Updown + rf + bs [25]	61.4	52.5	44.9	53.1	*11.8*	*18.2*	*36.5*	*35.7*
AoANet + rf + bs [4]	64.6	56.6	49.8	56.1	16.2	*20.3*	39.7	47.1
X-LAN + rf + bs [9]	**65.2**	**57.6**	**51.1**	**57.8**	**18.8**	**21.4**	**41.0**	**51.7**
X-Transformer + rf + bs [9]	63.9	56.2	*49.7*	57.0	17.4	*20.7*	40.4	48.5

Especially, the state-of-the-art image captioning models AoANet, X-LAN, and X-Transformer achieve a large improvement compared to the baseline system. X-LAN achieves the best performance, which is 39.6% and 90.1% relatively higher than the baseline R-I2P in terms on the BLEU4 metric on the generated phoneme sequences and text outputs respectively. Compared with the Image2Text model in [14], except for NIC [20], all other image captioning models obtain better performances.

C. The effect of reinforcement learning and beam search

Table II shows the performance of the different models with reinforcement learning and/or beam search. Similar to the previous section, we report the results using the different metrics on the image-to-phoneme output and on the image-to-text output. The results in Table II should be compared to those in Table I, which shows the models' performance without reinforcement learning and beam search. Looking at the image-to-phoneme output, we see that both reinforcement learning and beam search boost the performance of all image-to-phoneme models. Compared to the beam search, the reinforcement learning brings more improvements on the phoneme-based objective metric scores. The combination of the reinforcement learning and beam search brings a further slight improvement.

TABLE III: Comparison of the end-to-end SAS and speech unit based method and also the word-based image captioning method. Bold indicates the best result of each metric. ROUGE-L score was not reported by Katiyar et al.

Intermediate	Methods	BLEU4	METEOR	ROUGE-L	CIDEr
Word	Katiyar et al. [62]	**21.4**	20.0	—	**55.5**
Discovered Speech Unit	Hsu et al. [12]	12.5	14.5	39.1	24.5
	Effendi et al. [14]	14.8	17.4	**45.8**	32.9
Phoneme	R-I2P	8.1	15.7	30.3	21.7
	X-LAN	16.7	**21.4**	40.2	48.4
—	SAS	3.5	11.3	23.2	8.0

However, when we compare the text-based results in Table II and Table I, the improvements we observe of the phoneme-based results do not always lead to a better image-to-text performance. While the best image-to-text performance is achieved by the combination of reinforcement learning and beam search, only using reinforcement learning or in combination with beam search leads to an obvious decrease for the NIC, SAT, and Updown models. This decrease in performance when reinforcement learning is used could be caused by the reward scores, i.e., BLEU4, during reinforcement learning being calculated on the phoneme sequences with the aim to produce phoneme sequences with higher either BLEU4 score rather than being calculated on the word sequences. Interestingly, phoneme sequences with higher BLEU4 scores apparently do not necessarily lead to higher BLEU4 scores for the word sequences.

The use of beam search on the other hand seems to bring a relatively stable improvement on the image-to-text output. Note that there is no specific optimization on the phoneme sequences in term of objective metrics. Thus, the beam search could be an effective addition to the image-to-phoneme system.

VI. Results of image-to-speech

A. Objective Results

The baseline R-I2P and the best performing image-to-phoneme method, i.e., X-LAN, listed in Table I and Table II are compared with our SAS model on the image-to-speech task. In order to compare the output of the SAS model with those of the baseline and X-LAN models, we use the image-to-text output of the latter two models, while the synthesized speech of the SAS model was automatically transcribed into words using an ASR (see Section IV-D). Moreover, the SAS model's performance is compared to the two recently proposed speech unit-based methods [12], [14]. We also added a direct image-to-word (image captioning) model [62] to the table; this model shows state-of-the-art performance on Flickr8k, and can be taken as the upper-bound performance. The word-based captions are then evaluated in terms of BLEU4, METEOR, ROUGE-L, and CIDEr. The results are shown in Table III, with bold indicating the best performance for each metric.

Table III clearly shows that all the speech unit-based methods, i.e., the automatically discovered speech unit-based methods [62], [12] and the phoneme-based methods (the baseline and the X-LAN models) outperform our SAS method on all evaluation metrics. So, the performance of the end-to-end image-to-speech SAS model falls short of the performance of the image-to-speech models that use intermediate representations.

The explanation for the worse performance of the end-to-end SAS model is likely that the end-to-end image-to-speech task is much more challenging than the image-to-speech unit task, due to the following reasons: 1) for a given stretch of speech of the same duration, SAS generates a spectrogram sequence that is much longer than its transcribed phoneme sequence (because a phoneme consists of a sequence of spectra), while at the same time there is no explicit alignment between the spectrograms and image regions, making it impossible to explicitly learn the alignment as in [63], [64], [65], and 2) in the image-to-phoneme model, the phoneme generation process during inference can be seen as an autoregressive phoneme prediction process that predicts a phoneme based on an implicitly learned phoneme dictionary at each step. Consequently, it can generate a meaningful phoneme at each step, while there is no dictionary for spectra in the SAS model.

Regarding our question how well state-of-the-art image captioning models perform on the image-to-speech task, we can see that the phoneme-based X-LAN model outperforms the two best systems, i.e., the two automatically discovered speech unit-based methods, (except for the ROUGE-L metric, where Effendi et al. [14] outperforms the other models). Regarding the image-to-phoneme task, comparing the results of [14] (see also Table I) shows that X-LAN also outperforms [14] on the image-to-speech unit task. When X-LAN includes beam search (see Table II), the performance is even higher, showing that X-LAN performs well on the image-to-speech unit task.

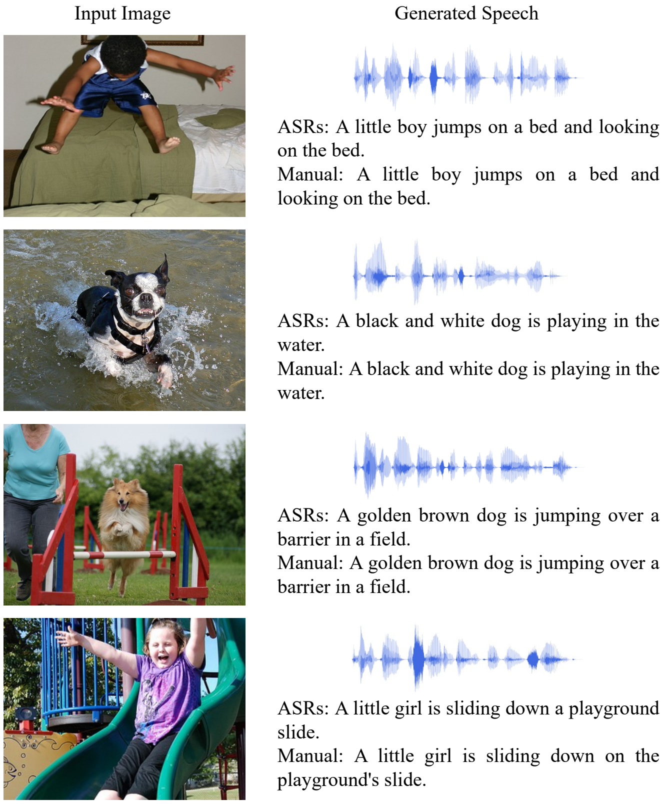
B. Visual inspection of some generated examples

Generated examples of some good and bad automatically generated spoken captions are shown in Fig. 3 and Fig. 4, respectively. For ease of the reader, the synthesized speech content was transcribed and presented below each image: "ASRs" means that the textual descriptions were created by the ASR system, and "Manual" means that the text was transcribed manually by a human listening to the synthesized speech without access to the corresponding images. The generated spoken captions of the examples in Fig. 3 and Fig. 4 and additional examples can be found on the project website[2].

As shown in Fig. 3, the proposed SAS model is able to generate spoken captions that describe the image well, indicating that end-to-end image-to-speech generation bypassing phonemes is feasible. Moreover, based on the fact that spoken captions of low audio quality would yield bad ASR transcriptions, the comparison of the transcriptions provided by the ASR system and those created by the human indirectly show the good quality of the synthesized speech.

However, there are also many cases where our SAS model failed to generate spoken captions that describe the image. Fig. 4 shows three such cases. In the top image, the synthesized

[2] https://xinshengwang.github.io/projects/SAS/



ASRs: A little boy jumps on a bed and looking on the bed.
Manual: A little boy jumps on a bed and looking on the bed.

ASRs: A black and white dog is playing in the water.
Manual: A black and white dog is playing in the water.

ASRs: A golden brown dog is jumping over a barrier in a field.
Manual: A golden brown dog is jumping over a barrier in a field.

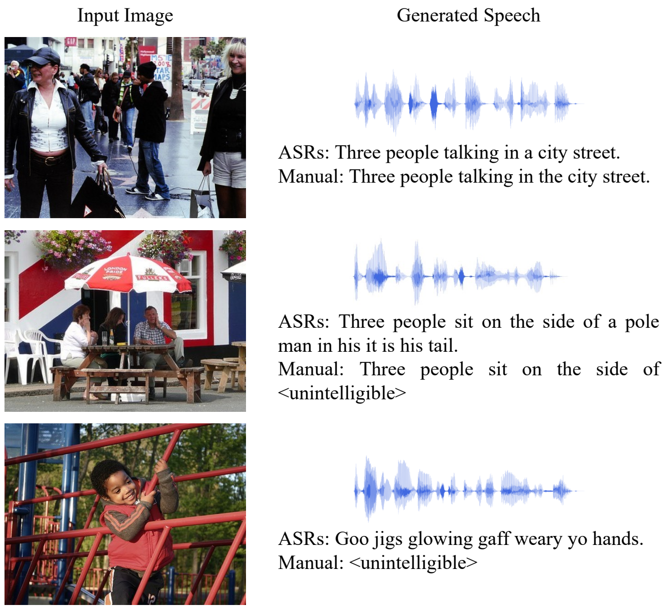
ASRs: A little girl is sliding down a playground slide.
Manual: A little girl is sliding down on the playground's slide.

Fig. 3: Examples of automatically generated spoken descriptions which describe the image well.

speech is of good quality, i.e., it is intelligible, but the spoken caption does not describe the image well. In the middle image, the quality of the synthesized speech is good at the beginning but gets worse throughout the spoken caption. The bottom image indicates the worst case in which the synthesized speech performance of the proposed method needs further improvement.

C. Further investigations into SAS

We further investigated the possibilities and limitations of our SAS model. In addition to the experiment performed on the Flickr8k database with synthesized spoken captions, two further experiments were carried out. First, we investigated the performance of the model when trained with multi-speaker natural speech for which the natural spoken captions from the 183 speakers from Flickr8k were used. These captions were collected by [47] using Amazon Mechanical Turk workers who were asked to pronounce the original written captions. Second, we investigated whether more training data would improve the performance of SAS for which an extended version of Flickr8k, i.e., Flickr30k [66] was used. Flickr30k consists of around 31000 images. Since no natural spoken captions exist for these 31000 images, all spoken captions in Flickr30k are synthetic speech. In all experiments, the same test set was used as for the first SAS experiment. In the multi-speaker Flickr8k experiment, the speaker id is represented by an embedded one-hot label that is concatenated with the encoder output of SAS. As the utterances from the training set and test set share the same speakers, during inference, we directly use the speaker id of each utterance.



ASRs: Three people talking in a city street.
Manual: Three people talking in the city street.

ASRs: Three people sit on the side of a pole man in his it is his tail.
Manual: Three people sit on the side of <unintelligible>

ASRs: Goo jigs glowing gaff weary yo hands.
Manual: <unintelligible>

Fig. 4: Examples of automatically generated spoken descriptions which do not describe the image well. The ⟨unintelligible⟩ in the manually transcribed text means that the corresponding speech is unintelligible.

TABLE IV: Performance of the SAS model with the natural multi-speaker spoken caption dataset and the larger training data Flickr30k. The results of our SAS model are repeated in this table for the convenience of the reader.

Dataset	BLEU4	METEOR	ROUGE-L	CIDEr
Flickr8k	3.5	11.3	23.2	8.0
Multi-speaker	0.4	6.8	15.7	0.7
Flickr30k	4.6	13.7	24.6	8.0

The results are shown in Table IV. For comparison, the performance on the synthetic single-speaker database of Flickr8k is also listed in this table (Flickr8k). As can be seen, the performance on the natural multi-speaker database is much worse compared to the performance on the synthetic single speaker database (Flickr8k). This phenomenon is easy to explain: 1) Speech synthesis models depend on high-quality recorded speech that is normally recorded by professional speakers in a recording studio. In contrast, the spoken captions of Flickr8k are recorded by amateurs with various styles and home equipment, which makes this database unsatisfactory for speech synthesis models; 2) Compared to single-speaker database, multi-speaker information also brings challenges to this task. In contrast, training with a larger database, i.e., Flickr30k, the performance improves substantially, although its performance still falls short of that of the two methods which use automatically discovered speech units and X-LAN.

D. Component analysis

As the image features showed an important impact on the image captioning task [25], the performance of the bottom-up features and vanilla ResNet features are compared in this

TABLE V: Results of the component analysis: The effect of image features, the image embedding constraint, and scheduled sampling on image-to-speech synthesis. Bold indicates the best result of each metric. The results of our SAS model are repeated in this table for the convenience of the reader.

Methods	BLEU4	METEOR	ROUGE-L	CIDEr
SAS	**3.5**	**11.3**	**23.2**	**8.0**
SAS-ResNet	3.1	11.0	22.4	7.3
SAS w/o EC	2.8	11.1	22.8	6.8
SAS w/o ss	2.7	10.8	22.8	6.7

section. Moreover, the effectiveness of the proposed image embedding constraint and the scheduled sampling during the training process are also investigated through an ablation study.

The results are shown in Table V, "SAS-ResNet" means the image features are extracted from the pre-trained ResNet-101 rather than the faster-RCNN, "SAS w/o EC" means that the SAS model drops the module of the image embedding constraint, and "SAS w/o ss" means that the SAS model is trained with the teacher-forcing method without using the scheduled sampling. As shown in the table, SAS shows better performance than SAS-ResNet, indicating that the bottom-up features outperform the ResNet-101 features in the image-to-speech task. The SAS w/o EC shows worse performance than the SAS on all metrics. Specifically, the BLEU4 score drops from 3.5 to 2.8, showing the importance of the image embedding constraint module. Training the model with the teacher-forcing method instead of the scheduled sampling also shows an obvious performance decrease, indicating the importance of the scheduled sampling strategy on training the end-to-end image-to-speech model.

VII. Discussion

Image-to-speech is a new task, closely related to the image captioning task, that tries to generate spoken descriptions of images, without the use of text as an intermediate representation. Previous work generated sequences of phonemes (which we here refer to as the image-to-phoneme task) that could be synthesized in a subsequent step (completing the image-to-speech task). In this work, 1) we compared different image captioning models and the effect of beam search and reinforcement learning on the image-to-phoneme task, and 2) compared different objective evaluation metrics to evaluate how well the generated phoneme sequences described the scenes in the images. Finally, 3) we presented an end-to-end image-to-speech method, bypassing any intermediate representation.

In order to find the best evaluation metric for the image-to-phoneme task, we compared several well-known objective metrics to several text-based objective metrics on the text outputs that were converted from the generated phoneme sequences. The comparisons of phoneme-based objective evaluation results with the human ratings in the previous work [15] and the transcribed text-based evaluation results showed that BLEU3, BLEU4, BLEU5, and ROUGE-L are most suitable to evaluate the image-to-phoneme task. Interestingly, all these metrics have medium length n-grams, i.e., 3-grams, 4-grams, 5-grams are used in BLEU3, BLEU4, and BLEU5 respectively, and the Longest Common Subsequence (LCS) is used in ROUGE-L.

The performance of the various image captioning models showed that image captioning models that are originally designed for the image-to-text task work reasonably well for the image-to-phoneme task, and better than the baseline model. The best models were X-LAN [9] and X-Transformer [9]. Their superior performance could be attributed to the X-Linear attention block, used in both models but not the baseline model, which can obtain a better interaction between different modalities, in this case speech and images.

The beam search in the inference process was found to have a positive effect on the performance of the different models on the image-to-phoneme task, both in terms of how well the generated phoneme sequences were able to describe the scenes in the images and in terms of the textual descriptions converted from the phoneme sequences. On the other hand, fine-tuning the image-to-phoneme models with the reinforcement learning brought improvements on the image-to-phoneme task, but failed to bring improvements at the text-output level. How to define a reward function that can consider the performance in terms of words could be an interesting topic for future research.

The experiments with the proposed end-to-end image-to-speech model SAS showed that synthesizing spoken descriptions of images bypassing any intermediate representation is feasible. The performance of the proposed model on the natural multi-speaker database is much worse than the performance on the synthetic single speaker database. This difference in performance is most likely due to the low quality of the recorded natural speech. However, generally, obtaining crowd-sourced data is a lot easier than obtaining high-quality speech recorded by professional speakers in a recording studio. This is typically especially the case for those languages that are low-resourced and/or unwritten. Since crowd-sourced data is of a lower quality than studio data, strategies need to be employed to deal with the noise and generally lower quality of these recordings. For instance speech enhancement [67] and background noise disentangling [68] could be considered for improving the performance of the end-to-end image-to-speech model trained with a low-quality natural multi-speaker database.

The current end-to-end model's performance is far behind those of the speech unit-based models, and the synthesized speech is not always of high quality or intelligible, which could be caused by the much longer frame sequence of the spectrogram and also the frame's variety. In order to improve on the image-to speech task, one could further improve the recently proposed approaches that use automatically discovered speech units by building better speech unit discovery models or use better image-to-speech unit models, e.g., X-LAN [9].

The motivation of the image-to-speech task is to allow speakers of unwritten languages benefit from text-independent image description technology. However, as unwritten languages typically are low-resourced, no existing appropriate databases, i.e., consisting of images and captions, of an actual unwritten language could be used for the current work. Future work should focus on moving beyond using English as the working language, and focus on building image-to-speech models for a real unwritten language.

VIII. Conclusion

This paper presents a study on synthesizing spoken captions of images, i.e., image-to-speech synthesis, bypassing the need for intermediate representations such as phonemes or text, and the evaluation of the generated captions. Extensive experiments demonstrate that standard image captioning models can be used in the image-to-phoneme task and automatic evaluation metrics, i.e., BLEU3, BLEU4, BLEU5, and ROUGE-L can be used to evaluate the performance of image-to-phoneme models. Additionally, the proposed end-to-end image-to-speech synthesis model, the first of its kind, showed that directly synthesizing spoken descriptions of images bypassing any text and phonemes is feasible, although its performance falls short of that of models that use phonemes as intermediate representation.

References

[1] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[2] S. Chen and Q. Zhao, "Boosted attention: Leveraging human attention for image captioning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 68–84.

[3] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8928–8937.

[4] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4634–4643.

[5] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 578–10 587.

[6] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7008–7024.

[7] T.-H. Chen, Y.-H. Liao, C.-Y. Chuang, W.-T. Hsu, J. Fu, and M. Sun, "Show, adapt and tell: Adversarial training of cross-domain image captioner," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 521–530.

[8] Y. Keneshloo, T. Shi, N. Ramakrishnan, and C. K. Reddy, "Deep reinforcement learning for sequence-to-sequence models," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 7, pp. 2469–2489, 2019.

[9] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 971–10 980.

[10] M. P. Lewis, G. F. Simons, and C. Fennig, "Ethnologue: Languages of the world (eighteenth edition)," 2015. [Online]. Available: http://www.ethnologue.com

[11] M. Hasegawa-Johnson, A. Black, L. Ondel, O. Scharenborg, and F. Ciannella, "Image2speech: Automatically generating audio descriptions of images," *Proceedings of ICNLSSP, Casablanca, Morocco*, 2017.

[12] W.-N. Hsu, D. Harwath, C. Song, and J. Glass, "Text-free image-to-speech synthesis using learned segmental units," *arXiv e-prints*, pp. arXiv–2012, 2020.

[13] D. Harwath, W.-N. Hsu, and J. Glass, "Learning hierarchical discrete linguistic units from visually-grounded speech," *arXiv preprint arXiv:1911.09602*, 2019.

[14] J. Effendi, S. Sakti, and S. Nakamura, "End-to-end image-to-speech generation for untranscribed unknown languages," *IEEE Access*, vol. 9, pp. 55 144–55 154, 2021.

[15] J. van der Hout, Z. D'Haese, M. Hasegawa-Johnson, and O. Scharenborg, "Evaluating automatically generated phoneme captions for images," in *INTERSPEECH 2020*. ISCA, 2020, pp. 2317–2321.

[16] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[17] R. Mason and E. Charniak, "Nonparametric method for data-driven image captioning," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2014, pp. 592–598.

[18] R. Socher and L. Fei-Fei, "Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 966–973.

[19] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu, "I2t: Image parsing to text description," *Proceedings of the IEEE*, vol. 98, no. 8, pp. 1485–1508, 2010.

[20] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[21] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[22] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 375–383.

[23] S. Bai and S. An, "A survey on automatic image caption generation," *Neurocomputing*, vol. 311, pp. 291–304, 2018.

[24] L. Gao, K. Fan, J. Song, X. Liu, X. Xu, and H. T. Shen, "Deliberate attention networks for image captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8320–8327.

[25] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[27] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," in *Advances in Neural Information Processing Systems*, 2019, pp. 11 137–11 147.

[28] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 649–665.

[29] D. Harwath, A. Torralba, and J. Glass, "Unsupervised learning of spoken language with visual context," in *Advances in Neural Information Processing Systems*, 2016, pp. 1858–1866.

[30] D. Merkx, S. L. Frank, and M. Ernestus, "Language learning using speech to image retrieval," *arXiv:1909.03795*, 2019.

[31] G. Ilharco, Y. Zhang, and J. Baldridge, "Large-scale representation learning from visually grounded untranscribed speech," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 2019, pp. 55–65.

[32] H. Kamper, G. Shakhnarovich, and K. Livescu, "Semantic speech retrieval with a visually grounded model of untranscribed speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 89–98, 2018.

[33] O. Scharenborg, L. Besacier, A. Black, M. Hasegawa-Johnson, F. Metze, G. Neubig, S. Stüker, P. Godard, M. Müller, L. Ondel *et al.*, "Speech technology for unwritten languages," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 964–975, 2020.

[34] X. Wang, T. Tian, J. Zhu, and O. Scharenborg, "Learning fine-grained semantics in spoken language using visual grounding," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2021, pp. 1–5.

[35] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," *International Journal of Computer Vision*, vol. 128, no. 3, pp. 620–641, 2020.

[36] L. Wang and M. Hasegawa-Johnson, "Multimodal word discovery and retrieval with spoken descriptions and visual concepts," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1560–1573, 2020.

[37] L. Wang, X. Wang, M. Hasegawa-Johnson, O. Scharenborg, and N. Dehak, "Align or attend? toward more efficient and accurate spoken word discovery using speech-to-image retrieval," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7603–7607.

[38] X. Wang, T. Qiao, J. Zhu, A. Hanjalic, and O. Scharenborg, "S2IGAN: Speech-to-Image Generation via Adversarial Learning," in *Proc. Interspeech 2020*, 2020, pp. 2292–2296. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1759

[39] ——, "Generating images from spoken descriptions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 850–865, 2021.

[40] G. Neubig, M. Sperber, X. Wang, M. Felix, A. Matthews, S. Padmanabhan, Y. Qi, D. S. Sachan, P. Arthur, P. Godard *et al.*, "Xnmt: The extensible neural machine translation toolkit," *arXiv preprint arXiv:1803.00188*, 2018.

[41] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[42] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7219–7228.

[43] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[44] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[45] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[46] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[47] D. Harwath and J. Glass, "Deep multimodal semantic embeddings for speech and images," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 237–244.

[48] K. Kilgour, M. Heck, M. Müller, M. Sperber, S. Stüker, and A. Waibel, "The 2014 KIT IWSLT speech-to-text systems for English, German and Italian," in *International Workshop on Spoken Language Translation (IWSLT)*, 2014, pp. 73–79.

[49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[51] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.

[52] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and VQA," in *AAAI*, 2020, pp. 13 041–13 049.

[53] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[54] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[55] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[56] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[57] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. INTERSPEECH 2018*, 2018, pp. 3743–3747.

[58] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[59] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 1171–1179.

[60] K. Ito, "The LJ speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[61] L. Guo, J. Liu, X. Zhu, P. Yao, S. Lu, and H. Lu, "Normalized and geometry-aware self-attention network for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 327–10 336.

[62] S. Katiyar and S. K. Borgohain, "Image captioning using deep stacked lstms, contextual word embeddings and data augmentation," *arXiv preprint arXiv:2102.11237*, 2021.

[63] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: fast, robust and controllable text to speech," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 3171–3180.

[64] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=piLPYqxtWuA

[65] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei *et al.*, "Durian: Duration informed attention network for speech synthesis." in *INTERSPEECH*, 2020, pp. 2027–2031.

[66] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[67] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech." in *SSW*, 2016, pp. 146–152.

[68] W.-N. Hsu, Y. Zhang, R. J. Weiss, Y.-A. Chung, Y. Wang, Y. Wu, and J. Glass, "Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5901–5905.