Show and speak

Directly synthesize spoken description of images

Wang, Xinsheng; Feng, Siyuan ; Zhu, Jihua; Hasegawa-Johnson, Mark; Scharenborg, Odette

DOI
10.1109/ICASSP39728.2021.9414021

Publication date
2021

Document Version
Final published version

Published in
ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414021

SHOW AND SPEAK: DIRECTLY SYNTHESIZE SPOKEN DESCRIPTION OF IMAGES

Xinsheng Wang[1,2†]*, Siyuan Feng*[2]*, Jihua Zhu*[1*]*, Mark Hasegawa-Johnson*[3]*, Odette Scharenborg*[2]

[1]School of Software Engineering, Xi'an Jiaotong University, Xi'an, China
[2]Multimedia Computing Group, Delft University of Technology, Delft, The Netherlands
[3]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, IL, USA
wangxinsheng@stu.xjtu.edu.cn, s.feng@tudelft.nl, zhujh@xjtu.edu.cn, jhasegaw@illinois.edu, o.e.scharenborg@tudelft.nl

ABSTRACT

This paper proposes a new model, referred to as the show and speak (SAS) model that, for the first time, is able to directly synthesize spoken descriptions of images, bypassing the need for any text or phonemes. The basic structure of SAS is an encoder-decoder architecture that takes an image as input and predicts the spectrogram of speech that describes this image. The final speech audio is obtained from the predicted spectrogram via WaveNet. Extensive experiments on the public benchmark database Flickr8k demonstrate that the proposed SAS is able to synthesize natural spoken descriptions for images, indicating that synthesizing spoken descriptions for images while bypassing text and phonemes is feasible.

Index Terms— Image-to-speech, image captioning, speech synthesis, sequence-to-sequence, encoder-decoder

1. INTRODUCTION

A system that can describe visual scenes in natural language has great potential for helping, for instance, visually-impaired people "see" the world. Recent research in this direction is called image captioning [1], which aims to automatically generate textual descriptions of images. Image captioning systems that automatically generate textual captions of images are inspired by the architecture of neural machine translation and have a neural encoder-decoder structure [2, 3]. Recently, benefiting from the development of attention mechanisms [4, 5, 6, 7, 8] and training strategies [9, 10, 11], the task of image captioning has achieved impressive results, making this technology more and more likely to be used in reality. However, nearly half of the world's languages do not have a written form [12], which means that speakers of those languages cannot benefit from any text-based technologies, including image captioning. In order to make this type of technology available for all languages, it is necessary to develop an image captioning method that bypasses text.

The first work that tried to generate image captions bypassing text is proposed by Hasegawa-Johnson et al. [13]. In their work, the authors proposed the image-to-speech task, which was based on an intermediate representation of the speech signal in terms of phoneme sequences. Their system first performs an image-to-phoneme generation process, after which the generated phoneme sequence can be used to synthesize the audio signal. Most recently, Van der Hout et al. [14] improved the image-to-phoneme part of the original system [13] by changing the image encoder structure. Moreover, they investigated how such an image-to-phoneme system could be evaluated objectively by comparing several objective evaluation measures to human ratings. Developing an image-to-phoneme system depends on large amounts of (automatic) alignments of the speech signal with the phonemes. Creating these phoneme annotations requires linguistic expertise. Although Hasegawa-Johnson et al. [13] also investigated the possibility of using automatically discovered speech units, those units performed subpar in the image-to-phoneme task, and the speech synthesis process based on such speech units was not investigated. Taken together, the image-to-phoneme-to-speech approach is difficult to implement for unwritten languages.

In order to make an image captioning system able to bypass the dependency on both text and phonemes, this paper presents an image-to-speech generation method which can synthesize spoken descriptions directly from images. The basic architecture of the proposed method is an attention-guided sequence-to-sequence model. Moreover, in order to suppress the embedding of image regions that would not be part of a human-generated description, an embedding constraint is implemented for the image encoder. This model, referred to as the Show and Speak (SAS) model, takes an image as input and outputs the synthesized spoken description of the image[1].

2. MODEL ARCHITECTURE

The proposed SAS model is designed as an encoder-decoder structure. Specifically, the encoder takes an image as input and outputs a sequence of feature vectors of this image. Then, these image feature vectors are taken as input to the decoder which then synthesizes the spectrogram of speech that describes the corresponding image. The architecture of the proposed method is shown in Fig. 1 and will be explained in detail below.

2.1. Encoder

The structure of the encoder is shown in the first column from left 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 [15], the Faster-RCNN [16] model pre-trained on ImageNet [17] and Visual Genome [18] is adopted to extract image features of $l = 36$ object regions, and these features are referred to as bottom-up features. 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 [15] provides its position in the image,

†Xinsheng Wang was supported by the China Scholarship Council (CSC).
*Corresponding author.

[1]The synthesized examples, source code, and database can be found at: https://xinshengwang.github.io/project/sas/

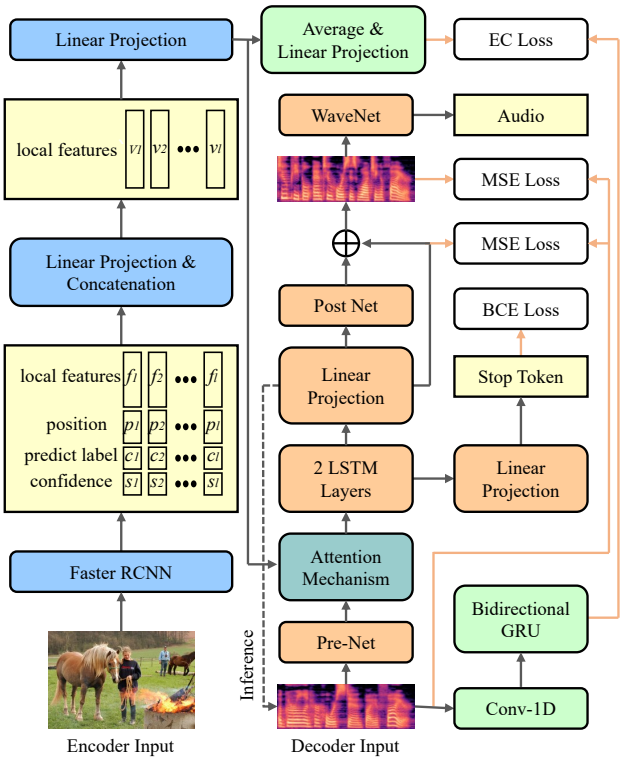

Fig. 1. Architecture of the Show and Speak (SAS) model.

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 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. Following [19], 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 the 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 transform 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 2.3.

2.2. Decoder

The structure of the decoder is shown in the middle column in 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 12.5 ms frame hop. The decoder architecture follows the structure of Tacotron2's decoder [20]. Specifically, the generated spectrogram frame from the previous time step passes through a Pre-Net and then is concatenated with an attention context vector before passing through two LSTM layers. The attention context vector is obtained from the encoder output with the location-sensitive attention [21], and the Pre-Net is a linear transform layer with 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 passes 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 an improved spectrogram 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 [22] in [20].

2.3. Objective function

Following the objective function in Tactron2 [20], mean squared error (MSE) is used to optimize the generation of spectrograms before and after Post-Net. Binary cross-entropy (BCE) is used to train the "Stop Token" prediction module that is similar to the module in [20]. The stop token prediction allows for the model to dynamically determine the length of the predicted spectrogram sequence instead of synthesizing a fixed-length sequence. 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 green background in Fig. 1 show the operations for the image embedding constraint. The image global feature vector is obtained by averaging the encoder outputs, and a linear transform 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 [23, 24]. Specifically, the ground-truth speech spectrogram first passes through a 1-D convolutional layer, and the fixed-length speech feature vector 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 [25] to obtain the EC loss. We denote losses of spectrogram synthesis, stop token prediction, and image embedding constraint by $\mathcal{L}_s$, $\mathcal{L}_{st}$, and $\mathcal{L}_{ec}$ respectively. The total loss for training the SAS model in an end-to-end way is given by

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_{st} + \lambda\mathcal{L}_{ec}, \tag{2}$$

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

3. EXPERIMENTAL SETTING

3.1. Database

Following the previous experiments on the image-to-phoneme task [13, 14], Flickr8k [26] is used in our experiments. This database contains 8,000 images, and each image has 5 textual descriptions. There is a Flickr-Audio database [27] which contains speech recordings for the corresponding textual descriptions. However, these recordings come from more than 100 different speakers, making speech

synthesis quite a challenging task. Here, we will not take the multi-speaker speech synthesis problem into consideration, but adopt a text-to-speech (TTS) system [20] to synthesize the spoken captions on the basis of textual captions from a single speaker. This TTS system is pre-trained on LJSpeech [28] which consists of 13,100 audio clips recorded from a single speaker. We split Flickr8k in the standard way: 6,000 images for training and 1,000 images both for development and test set.

3.2. 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. In order to objectively measure the performance of our system, the synthesized speech is automatically transcribed to text. To that end, an automatic speech recognition (ASR) system[2] built with Kaldi [29] is adopted. The ASR system consists of a hybrid factorized time-delay neural network (TDNN-F) [30] acoustic model (AM) and a four-gram language model (LM), both trained using the 960-hour Librispeech English database [31].

The transcribed textual captions are then evaluated using evaluation metrics for image captioning [7, 8]: BLEU1 (B1), BLEU2 (B2), BLEU3 (B3), BLEU4 (B4), METEOR (M), ROUGE (R), and CIDEr (C). Because the evaluation is performed on the textual captions that are transcribed from the speech captions via the ASR system, higher scores of those metrics can also reflect the good quality of synthesized speech to a certain extent as a worse quality of the synthesized speech would seriously affect the accuracy of the ASR system.

3.3. 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 [32] to alleviate this problem. However, we found that when the percent of ground-truth input during the training process decreases to a small value, the generation of speech would be seriously affected. So, we use the inverse sigmoid decay method [32] for the percent of ground-truth input with a minimum value of 97.5%. The effect of this minimum value will be discussed in Section 4.3.

4. RESULTS

4.1. Objective Results

As the proposed SAS model is the first method that directly synthesizes spoken descriptions of images, there are no existing methods to which it can be compared fairly. Therefore, we take the state-of-the-art image-to-phoneme method [14] to present an upper bound performance on the image-to-speech task. Since this method [14] only generates phonemes rather than speech, we first synthesized the speech based on the generated results of [14]. Specifically, we synthesized the speech with the word sequences generated by [14] (which were used in their human rating study) using the same TTS system that we used to create our ground-truth speech data [20]. This topline model is referred to as the phoneme-based method from here onwards.

Note that the image-to-phoneme system [14] is based on phonemes that were obtained using a well-trained same-language ASR, which means that the phoneme-based method cannot be applicable to an unwritten language (see also the Introduction).

Table 1 shows that the phoneme-based method outperforms our SAS method on all evaluation metrics. The explanation for the worse performance of our SAS model is likely that the end-to-end image-to-speech task is much more challenging than the image-to-phoneme task, due to the following reasons: 1) for a stretch of speech, SAS's spectrogram sequence is much longer than its transcribed phoneme sequence, 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 spectrograms in the SAS model.

Table 1. Compared with the phoneme-based (image-to-phoneme-to-speech) method.

	B1	B2	B3	B4	M	R	C
Phoneme-based	47.0	28.5	16.6	9.9	16.7	33.3	23.7
SAS	29.6	14.7	7.2	3.5	11.3	23.2	8.0

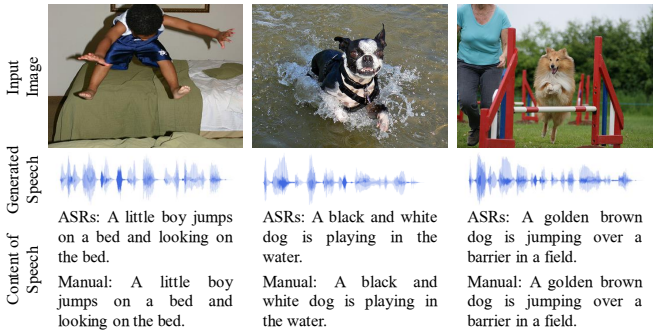

Fig. 2. Examples of good synthesized spoken descriptions.

Fig. 3. Examples of bad synthesized spoken descriptions. The ⟨unintelligible⟩ in the manually transcribed text means that the corresponding speech is unintelligible.

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

4.2. Subjective Results

Subjective results that display some good and bad automatically generated spoken captions are shown in Fig. 2 and Fig. 3 respectively. To show the speech content, transcribed textual descriptions from the synthesized speech are presented below each image: "ASRs" means the textual descriptions are given by the ASR system, and "Manual" means the corresponding text is transcribed manually by a human listening to the synthesized speech who does not have access to the corresponding images. The corresponding generated spoken captions and additional examples can be found on the project website[1].

As shown in Fig. 2, the proposed SAS model can correctly synthesize spoken descriptions for these images, 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 synthesize good spoken captions. Fig. 3 shows three such cases. In the left image, the synthesized 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 right image indicates the worst case in which the synthesized speech is unintelligible. These cases indicate that the robustness of the proposed method needs further improvement.

4.3. Component analysis

Because the image features showed an important impact on the image captioning task [15], the performance of the bottom-up features and vanilla ResNet features are compared in this section. Moreover, the proposed image embedding constraint assisting the image embedding is investigated through an ablation study. Finally, the minimum percentage of guided input in the scheduled sampling training process is discussed.

The effect of the image features and the image embedding constraint is shown in Table 2. In this table, the Baseline is based on the image features extracted from the pre-trained ResNet-101 rather than the faster-RCNN. Specifically, the image vectors $v_1, v_2, ..., v_l$ in Fig. 1 are created by scanning the last convolutional layer of ResNet-101 in raster-scan order. In the Baseline, the image embedding constraint is not included. SAS w/o EC means that the SAS model drops the module of image embedding constraint. Compared to the Baseline, in SAS w/o EC, the ResNet-101 features are replaced by the bottom-up features as introduced in Section 2.1. As shown in this table, the SAS w/o EC shows better performance than the Baseline, indicating the bottom-up features outperform the ResNet-101 features in the image-to-speech task. When the proposed image embedding constraint is added (SAS), the performance increases further, showing the good performance and effectiveness of the image embedding constraint module.

Table 2. The effect of image features and the image embedding constrain on the image-to-speech synthesis. w/o means without.

Method	B1	B2	B3	B4	M	R	C
Baseline	28.8	13.1	5.6	2.5	10.4	22.0	5.5
SAS w/o EC	29.6	13.9	6.3	2.8	11.1	22.8	6.8
SAS	**29.6**	**14.7**	**7.2**	**3.5**	**11.3**	**23.2**	**8.0**

The effect of the minimum percent ε of guided sampling (see Section 3.3) during the scheduled sampling training process is shown in Table 3. In this table, $\varepsilon = 100$ means the model is trained fully in a teacher-forcing way, and $\varepsilon = 90$ means the percent of ground-truth input exponentially decreases to 90% from 100% during the training. As shown in this table, when the scheduled sampling strategy is implemented ($\varepsilon < 100$), the performance shows obvious changes compared to the fully teacher-forcing training method. Specifically, when the minimum percent of ground-truth input is set as 97.5%, the SAS model shows the best performance which achieves 29.6% relative improvement on the BLEU4 compared to the fully-forcing training method. However, when the sampling rate from the real spectrograms (ground-truth input) goes too low, the scheduled sampling leads to a negative effect on the results. Specifically, when $\varepsilon < 95.0$, the generated results become worse than the teacher-forcing training ($\varepsilon = 100$).

Table 3. The effect of minimum guided sampling rate on training the SAS model.

ε	B1	B2	B3	B4	M	R	C
100.0	30.0	13.7	5.8	2.7	10.8	22.8	6.7
99.0	**30.2**	14.4	6.5	3.1	11.1	22.5	7.0
97.5	29.6	**14.7**	**7.2**	**3.5**	**11.3**	**23.2**	**8.0**
95.0	28.5	13.6	6.2	3.1	10.8	22.3	7.0
92.5	27.8	13.0	6.1	2.8	10.9	21.6	6.3
90.0	25.8	12.3	5.8	2.7	10.0	20.7	6.6

5. DISCUSSION AND CONCLUSION

This paper proposes an image-to-speech model, named SAS, which, for the first time, can generate spoken captions of images directly, bypassing any text and phonemes. The proposed SAS model takes an image as input and outputs a spoken description of the image. The results of the image-to-speech experiments show that our SAS model can indeed generate natural spoken descriptions that correctly describe the images.

Although in the phoneme-based method [13], the authors have tried to use automatically discovered speech units as intermediaries (to replace the phonemes), no evidence has shown that these automatically discovered speech units can be used to synthesize natural speech (the speech synthesis stage was not implemented). Moreover, the automatically discovered speech unit-based method showed much worse performance compared to the phoneme-based method on the image-to-speech unit generation task (i.e., without considering the synthesis stage of speech). Performance of the image-to-phone system is better if it uses phonemes transcribed by a well-trained same-language ASR, but as stated by the authors in [13], such a system cannot be used for unwritten languages. The proposed SAS model is the first method that can be used to synthesize natural speech to describe images for unwritten languages, and builds a baseline for this task.

Compared to the upper bound given by the phoneme-based method [14], the proposed SAS model still has a large gap to bridge. The SAS model does not always synthesize intelligible speech. An adversarial learning strategy could be considered in the future to improve the quality of the synthesized speech. Finally, the current work is based on the well-resourced English language, and it will be highly interesting to implement this task on a real unwritten language in the future.

6. REFERENCES

[1] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys (CSUR)*, vol. 51, no. 6, pp. 1–36, 2019.

[2] 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.

[3] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2407–2415.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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*, 2019.

[12] M. P. Lewis, G. F. Simons, and C. Fennig, "Ethnologue: Languages of the world [eighteenth," *Dallas, Texas: SIL International*, 2015.

[13] 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.

[14] J. van der Hout, Z. D'Haese, M. Hasegawa-Johnson, and O. Scharenborg, "Evaluating automatically generated phoneme captions for images," in *Proc. INTERSPEECH 2020*, 2020.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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 *ICASSP*. IEEE, 2018.

[21] 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.

[22] 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.

[23] 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.

[24] X. Wang, T. Qiao, J. Zhu, A. Hanjalic, and O. Scharenborg, "Generating images from spoken descriptions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[25] 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.

[26] 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.

[27] 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.

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

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[30] 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.

[31] 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.

[32] 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.