Image2speech: Automatically generating audio descriptions of images

Mark Hasegawa-Johnson[1], *Alan Black*[2], *Lucas Ondel*[3], *Odette Scharenborg*[4], *Francesco Ciannella*[2]

1. University of Illinois, Urbana, IL USA 2. Carnegie-Mellon University, Pittsburgh, PA USA
3. Brno University of Technology, Brno, Czech Republic
4. Centre for Language Studies, Radboud University, Nijmegen, Netherlands

Abstract

This paper proposes a new task for artificial intelligence. The image2speech task generates a spoken description of an image. We present baseline experiments in which the neural net used is a sequence-to-sequence model with attention, and the speech synthesizer is clustergen. Speech is generated from four different types of segmentations: two that require a language with known orthography (words and first-language phones), and two that do not (pseudo-phones and second-language phones). BLEU scores and token error rates indicate that the task can be performed with better than chance accuracy. Informal perusal of the output (phone strings, word strings, and synthesized audio) suggests that the audio contains complete, intelligible words organized into intelligible sentences, and that the most salient errors are caused by mis-recognition of objects and actions in the image.[1]

1. Introduction

This paper proposes a new task for artificial intelligence: the generation of a spoken description of an image. The automatic generation of text is the topic of natural language processing (NLP), whereas the analysis of images is the topic of the field of computer vision. In both fields, great advances have been made on these separate topics, and recently they have been combined into a new research field: img2txt [1]. However, many of the world's languages do not have a written form [2], therefore many people do not have access to these and other speech and NLP technologies. In this work, we propose a new research task: image2speech, which is similar to img2txt, but can reach people whose language does not have a natural or easily used written form. An image2speech system should generate a spoken description of an image directly, without first generating text.

Experiments reported in this paper convert image feature vectors into speech unit sequences. In order to implement this pipeline, four types of standard open-source software toolkits are used. First, the VGG16 [3, 4] visual object recognizer converts each image into a sequence of feature vectors. Second, the XNMT [5] machine translation toolkit accepts image feature vectors as inputs, and generates speech units as output. Third, the ClusterGen [6] speech synthesis toolkit generates audio from each speech unit sequence. Fourth, in order to train a synthetic speech voice, Clustergen needs a corpus of audio files, each of which is transcribed using some type of discrete symbolic units; automatic speech recognition (ASR) systems based on Kaldi [7] and Eesen [8] perform this transcription.

The complete image2speech system is trained using a corpus of (image,description) pairs, where each description is an audio file. Four different types of speech units are tested, distinguished by the type of technology used to segment the audio training data. Two types of unit sequences, Words and L1-Phones (first-language phones), are generated using a same-language ASR, and would therefore never be applicable to a language without orthography, but they provide us with an upperbound performance on the image2speech task. Two other unit sequences, L2-Phones and Pseudo-Phones, are generated without transcribed same-language speech, and would therefore be applicable even in a language lacking orthography. L2-phones (second-language phones) are generated by an ASR that has been trained in some other language. Pseudo-phones are generated by an unsupervised acoustic unit discovery system.

This paper describes preliminary experiments in the image2speech task. Section 2 describes toolkits and baseline methods. Section 3.1 describes datasets. Section 3 describes methods. Section 4 presents numerical results for two img2txt baselines, and four image2speech experimental systems. Section 5 gives example image2speech outputs. Section 6 concludes.

2. Background

Imagenet [9] is an image database organized according to the WordNet [10] noun hierarchy. ImageNet currently has 14m images, provided as examples of 22k nouns. The ILSVRC (Imagenet Large Scale Visual Recognition Challenge) has been held annually since 2010. The best single-network solution in ILSVRC 2014 Sub-task 2a, "Classification+localization with provided training data," was a 13-layer convolutional neural network (CNN) [3]; Implementations in TensorFlow ([4], used in this paper) and Keras [11] are now redistributed as the VGG16 network. VGG16 is a 13-layer CNN, followed by a two-layer fully-connected network (FCN). The last convolutional layer is composed of 512 channels, each of which is a 14×14 image; it is useful to interpret this layer as a set of $14 \times 14 = 196$ feature vectors of dimension 512. Each feature vector is the nonlinear transformation of a 40×40-pixel sub-image, which is to say, about 3% of the original 224×224 input-image.

XNMT (the eXtensible Machine Translation Toolkit [5]) was used to implement/train the image2speech system. XNMT is specialized in the training of sequence-to-sequence neural networks, which means it reads in a sequence of inputs, and then generates a different sequence of outputs.

XNMT is based on DyNet [12], a library for the training of neural networks with variable-length inputs. Prior to DyNet, most neural network modeling toolkits assumed that every train-

[1] This work was started at the Jelinek Speech and Languate Technology Workshop 2017, in Pittsburgh, supported by JHU and CMU via grants from Google, Microsoft, Amazon, Facebook, and Apple. The Extreme Science and Engineering Discovery Environment (XSEDE) is supported by National Science Foundation grant number ACI-1548562. O.S. was partially supported by a Vidi-grant from NWO (276-89-003).

ing and test input is exactly the same size. DyNet introduced a new type of graph compilation: dynamic compilation, in which each layer of the neural net is represented as a compiled function, rather than a compiled data structure.

XNMT [5] is a DyNet-based library of standard components frequently re-used in neural machine translation. The library is designed so that existing components can be easily rearranged to run new experiments, and new components can be easily added. Available components are categorized as embedders (e.g., one-hot, linear, and continuous vector embedders), encoders (e.g., CNN, LSTM and pyramidal LSTM encoders), attention models (e.g., dot product, bilinear, and MLP attention models), decoders (e.g., an MLP decoder applied to the state vector of the encoder), and error metrics (e.g., BLEU, cross-entropy, word error rate). Among other applications, the flexibility of XNMT has been demonstrated in the use of attention models to select between neural and phrase-based translation probability vectors, a method that has particular utility in the translation of low-frequency content words [13].

Text-to-speech synthesis is typically a four-stage process. First, the text is converted to a graph of symbolic phonetic descriptors. Second, the duration of each unit in the phonetic graph is predicted. Third, the mel-cepstrum [14], pitch, and multi-band excitation [15] are predicted using a dynamic model such as an HMM (hidden Markov model, [16]) or RNN (recurrent neural network, [17]), or by applying separate discrete-to-continuous mapping algorithms to each frame of the synthetic utterance [6]. Fourth, the speech signal is generated by inverting the mel-cepstral transform [14], and exciting it with the specified excitation.

The Clustergen speech synthesis algorithm [6] differs from most other speech synthesis algorithms in that there is no predetermined set of speech units, and there is no explicit dynamic model. Instead, every frame in the training database is viewed as an independent exemplar of a mapping from discrete inputs to continuous outputs, and a machine learning algorithm (e.g., regression tree [6] or random forest [18]) is applied to learn the mapping. Clustergen is particularly applicable to the problems considered in this paper because it is able to generate intelligible and pleasant synthetic voices from very small training corpora, and using an arbitrary discrete labeling of the corpus that need not include any traditional type of phoneme [19].

3. Experimental Methods

Fig. 1 gives an overview of experimental methods used in this paper. image2speech models were trained using (image,audio) pairs drawn from the Flickr8k, MSCOCO, Flicker-Audio, and SPEECH-COCO corpora. Each image is represented as a sequence of 196 vectors, each of dimension 512, created from the last convolutional layer of the VGG16 network. Audio files are converted to units via Kaldi forced alignment (Words and L1-phones) or via Eesen or AMDTK phone recognition (L2-phones and Pseudo-phones). XNMT then learns to convert a sequence of image feature vectors into a sequence of speech units, while Clustergen learns to convert speech units into audio.

The image2speech model learned by XNMT is a sequence-to-sequence model, composed of an encoder, an attender, and a decoder. The encoder is a one-layer bidirectional LSTM (implemented using XNMT's PyramidalLSTM model), with a 128-dimensional state vector. The attender is a three-layer perceptron, implemented using XNMT's StandardAttender model. For each combination of an input LSTM state vector and an output LSTM state vector (128 dimensions each), the attender uses a three-layer perceptron (two hidden layers of 128 nodes each) to compute a similarity score. The decoder is another three-layer perceptron (1024 nodes per hidden layer), which views an input created as the attention-weighted summation of all input LSTM state vectors, concatenated to the state vector of the output LSTM. The output of the decoder is a softmax with a number of output nodes equal to the size of the speech unit vocabulary.

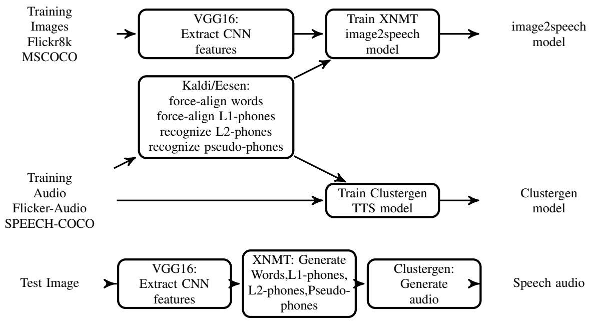


Figure 1: Experimental methods. XNMT and Clustergen models are first trained using (image,audio) pairs. A test image is then passed through VGG16, XNMT, and Clustergen to generate its audio description.

3.1. Data

Experiments in this paper used two databases: the Flickr8k image captioning corpus with its associated Flicker-Audio speech corpus, and the MSCOCO image captioning corpus with its associated SPEECH-COCO speech corpus.

The Flickr8k Corpus [20] includes five text captions for each of 8000 images, as well as links to the images. Text captions were written by crowd workers, hired on Amazon Mechanical Turk. There is considerable variability among the captions provided for each image. For example, the five different captions available for the first image in the corpus (image 1000268201_693b08cb0e) are:

A child in a pink dress is climbing up a set of stairs in an entry way.
A girl going into a wooden building.
A little girl climbing into a wooden playhouse.
A little girl climbing the stairs to her playhouse.
A little girl in a pink dress going into a wooden cabin.

In 2015, Harwath and Glass [21] proposed a data retrieval system in which speech files are used to retrieve corresponding images from a large image database, and vice versa. In order to make their proposal possible, they hired crowd workers on Mechanical Turk to read aloud the 40,000 captions from the Flickr8k corpus. The resulting set of 40,000 spoken captions is distributed as the Flicker-Audio corpus.

The Microsoft COCO (Common Objects in COntext) corpus was initially developed as an object detection corpus [22]. After initial release of the corpus, text captions of 150,000 of the images (four captions each) were distributed [23], making MSCOCO the largest database available for training img2txt systems. The SPEECH-COCO spoken transcriptions [24] were created using eight different synthetic voices, reading the MSCOCO text transcriptions. All eight synthetic voices were created from low-noise recordings of professional broadcast announcers; most listeners can't tell that the speech is synthetic. Because the speech is synthetic, exact time alignment of the phones and words is available, and is distributed with the corpus.

Experiments in this paper did not use the entire SPEECH-COCO corpus, because we did not have enough compute time to train a neural network using the whole corpus. Instead, experiments reported in this corpus used a subset of MSCOCO with training, validation, and test corpora sized to match those of Flickr8k: 6000 training images, 1000 validation images, and 1000 test images. When an image is part of the training or validation corpus, all of its captions are used, thus experiments using the MSCOCO corpus had a training corpus containing 24,000 image-audio pairs (6000 distinct images), while the Flickr8k training corpus included 30,000 image-audio pairs (6000 distinct images).

3.2. Speech Units

Systems were trained and tested using four different types of speech units: Words, L1-Phones, L2-Phones, and Pseudo-Phones.

Words and L1-Phones are aligned to the speech2image training audio files using ASR forced alignment trained in the target language, therefore speech segmentations of this kind can only be performed in a language that has a writing system. The two databases used in this paper were transcribed in two different ways. The larger corpus, SPEECH-COCO [24], is distributed with phone transcriptions (phones in this database are transcribed using a phonetic alphabet based on X-SAMPA). The Flicker-Audio corpus [21] is not distributed with phonetic transcriptions, but text transcriptions are available [20]; from these, aligned L1-Phone transcriptions were generated using the KIT English transcription system [25].

L2-Phone transcription does not use any information about the writing system of the target language, and could therefore be used in a language that lacks any writing system. In this method, an ASR is first trained in a different language (in our case, Dutch). The L2 ASR is then used to generate a phone transcription of the target audio. In the experiments reported in this paper, an Eesen speech recognizer [8] was first used to train a Dutch ASR. Dutch phones were then mapped to English phones using linguistic knowledge only, without the use of any English writing or transcriptions, and the English-adapted Dutch ASR was used to transcribe English audio. Thus the ASR has some prior exposure to English audio, but has no knowledge about English text [26].

Pseudo-Phones were generated from the Acoustic Unit Discovery (AUD) system of [27] with two major modifications. First, the truncated Dirichlet process of [27] was replaced by a symmetric Dirichlet distribution, since, as pointed out in [28], the symmetric Dirichlet distribution provides a good and yet simple approximation of the Dirichlet Process. Second, to cope with the relatively large database, the Variational Bayes Inference algorithm originally used in [27] was replaced with the faster Stochastic Variational Bayes Inference algorithm. It was found experimentally that these modifications, while considerably speeding up the training, yield negligible drop in accuracy. The source code of the AUD model is available at
https://github.com/amdtkdev/amdtk.

4. Results

The baseline models, with Word-sequence outputs, are standard img2txt networks, e.g., comparable to the result reported in [20]. In these networks, the output vocabulary of the network is the set of all distinct words in the training corpus: 7993 words in Flickr8k, 7476 in MSCOCO8k.

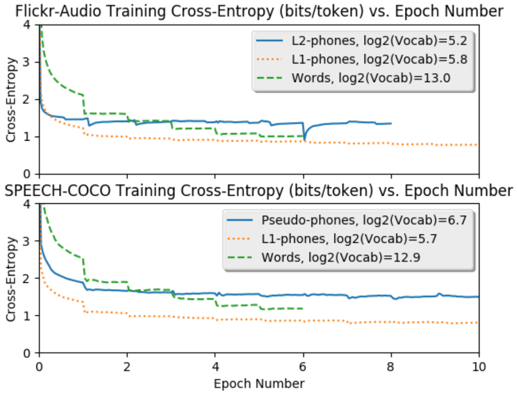


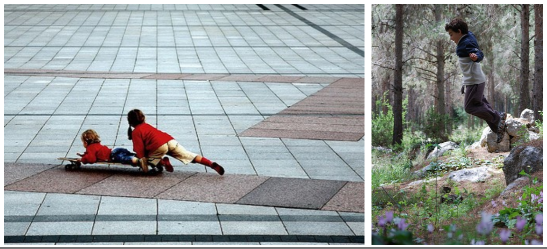
Figure 2: Training loss (cross-entropy, bits/symbol) of sequence-to-sequence networks trained to generate speech unit sequences from image features.

Table 1: BLEU scores (%) and unit error rates (UER, %) achieved in two baseline img2txt experiments (Word outputs) and four experimental image2speech experiments (L1-Phone, L2-Phone, and Pseudo-Phone), on both validation and test sets. ** means UER < chance (Student's T, Chebyshev standard error, $p < 0.001$; chance=90.2% Flickr8k, 88.7% MSCOCO).

	Validation		Test	
Dataset, Targets	BLEU	UER	BLEU	UER
Flickr8k, Words	4.7%	91.3%	3.7%	130%
Flickr8k, L1-Phones	13.7	87.9	13.7	84.9**
Flickr8k, L2-Phones	5.4	115	6.1	101
MSCOCO, Words	4.8		5.5	88.5
MSCOCO, L1-Phones	15.1		16.3	78.8**
MSCOCO, Pseudo-Ph.	2.2		1.4	123

The experimental systems generate phone outputs: L1-Phone, L2-Phone, and Pseudo-Phone. Flickr8k and MSCOCO L1-phone systems both use English phone sets, but with slightly different sizes: 54 for Flickr8k, 52 for MSCOCO. The L2-Phone system (only tested for Flickr8k) contains 38 phones. The Pseudo-Phone system (only tested for MSCOCO) was adjusted to produce 103 phones, as pseudo-phone sets of about this size have been useful in previous experiments [27].

Fig. 2 shows the training loss (cross-entropy) of sequence-to-sequence networks trained (using XNMT) to generate speech unit sequences from image features. Training loss is measured in bits per output symbol. Word-generating models start with training loss much higher than that of any phone-generating model, apparently because the number of distinct words is larger than the number of distinct phones. Training loss of the Word networks falls below those of the L2-phone and Pseudo-phone networks after some training, apparently because Words are more predictable than Pseudo-Phones or L2-Phones. The Word models never achieve training losses below those of the L1-Phone models. In fact, the Word and L1-Phone models converge to very similar endpoints, suggesting that the L1-Phone network may be learning the same thing as the Word model: it might be learning to generate a sequence of phones that always corresponds to complete Words.



Flickr8k Example #1
Ref #1: The boy +um+ laying face down on a skateboard is being pushed along the ground by +laugh+ another boy.
Ref# 2: Two girls +um+ play on a skateboard +breath+ in a court +laugh+ yard.
Network: SIL +BREATH+ SIL T UW M EH N AA R R AY D IX NG AX R EH D AE N W AY T SIL R EY S SIL.

Flickr8k Example #2
Ref #1: A boy +laugh+ in a blue top +laugh+ is jumping off some rocks in the woods.
Ref #2: A boy +um+ jumps off a tan rock.
Network: SIL +BREATH+ SIL EY M AE N IH Z JH AH M P IX NG IH N DH AX F AO R EH S T SIL.

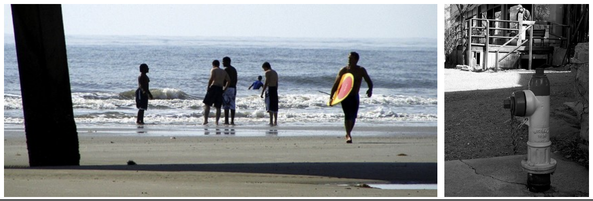
Figure 3: Image examples from the flickr8k corpus. The table lists, for each image, two of its reference transcriptions, and the output of the L1-Phone image2speech system.

Table 1 shows BLEU scores (higher is better) and unit error rates (UER; lower is better) of four experimental systems and two baselines, measured on the validation and test sets of the Flicker-Audio and MSCOCO8k corpora. For Word-generating systems, UER=word error rate; for Phone-generating systems, UER=phone error rate. Rank-ordering of the experimental systems is roughly the same in Table 1 as in Fig. 2, though the Word-based system achieves a very poor unit error rate on the Flickr8k test corpus. Both the L2-Phone and Pseudo-Phone systems suffer UER$>$ 100%. The L1-Phone systems, however, demonstrate unit error rates that are significantly better than chance (where "chance" is the error rate of a system that always generates the majority phone label).

Synthetic speech examples have been generated by the Clustergen algorithm for some of the L1-phone network outputs. Quality of the audio examples has not yet been quantified, but informal listening confirms the impression given by Fig. 2: generated audio is not perfectly natural, but is composed of intelligible words arranged into intelligible sentences.

5. Examples

Fig. 3 shows examples of two images from the validation subset of the Flickr8k corpus. For each image, three transcriptions are shown: two of the five available reference transcriptions (to give the reader a feeling for the difference among reference transcriptions), and one transcription generated by the L1-Phone image2speech network. The L1-Phones for Flickr8k are the ARPABET phones of [25]. As shown, the network is able to generate a phone string that is composed entirely of intelligible words, sequenced in an intelligible and semantically reasonable sentence. In these two examples, the phone strings shown can be read as English sentences that mislabel boys as men, but are otherwise almost plausible descriptions of the images: "Two men are riding a red and white race," and "A man is jumping in the forest."



MSCOCO Example #1
Ref #1: A group of men enjoying the beach, standing in the waves or surfing.
Ref# 2: A group of people standing on a beach next to the ocean.
Network: # @ g r uu p @ v p ii p l= s t a n d i ng o n @ b ii ch #

MSCOCO Example #2
Ref #1: A, a black and white photo of a fire hydrant near a building.
Ref #2: Aa, a fire hydrant that is out next to a house.
Network: # @ p @@ s n= w oo k i ng @ t @ m e dÎ= d au n @ n d @ r e d f ai r h ai d r @ n t #.

Figure 4: Image examples from the MSCOCO corpus. The table lists, for each image, two of its reference transcriptions, and the output of the L1-Phone image2speech system.

Fig. 4 shows similar examples from the SPEECH-COCO corpus. In the first example, the network has generated the sentence "A group of people standing on a beach," which is perfectly correct. In the second example, the network generated "A person working at a metal down, and a red fire hydrant." It is interesting that the neural net has noticed something about the image (the person working in the background) that was not noticed by either of the human transcribers.

6. Conclusions

This paper proposes a new task for artificial intelligence: image2speech, the task of generating spoken descriptions of input images, with no intermediate text. image2speech is trained using a database of paired images and audio descriptions. Experimental results are presented using the Flicker-Audio and SPEECH-COCO corpora. Measured UER scores are better than chance, but less than perfect. Informal perusal of results shows that image2speech is able to generate intelligible words, and to sequence them into intelligible sentences.

7. References

[1] J-Y Pan, H-J Yang, P Duygulu, and C Faloutsos, "Automatic image captioning," in *Proc. IEEE Internat. Conf. Multimedia and Expo (ICME)*, 2004.

[2] G Adda, S Stüker, M Adda-Decker, O Ambouroue, L Besacier, D Blachon, M Bonneau-Maynard, P Godard, F Hamlaoui, D Idiatov, G-N Kouarata, L Lamel, E-M Makasso, A Rialland, M Van de Velde, F Yvon, and S Zerbian, "Breaking the unwritten language barrier: The BULB project," in *Proceedings of the SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages*, 2016.

[3] K Simonyan and A Zisserman, "Very deep convolutional networks for large-scale image classification,"

https://arxiv.org/abs/1409.1556, 2014, Accessed: 2017-09-14.

[4] D Frossard, "Vgg16 in tensorflow," https://www.cs.toronto.edu/ frossard/post/vgg16/, 2016, Accessed: 2017-09-14.

[5] G Neubig, "eXtensible Neural Machine Translation," https://github.com/neulab/xnmt, 2017, Accessed: 2017-09-14.

[6] AW Black, "CLUSTERGEN: A statistical parametric speech synthesizer using trajectory modeling," in *Proc. Internat. Conf. Spoken Language Process. (ICSLP)*, 2006, pp. 1762–1765.

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

[8] Y Miao, M Gowayyed, and F Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015.

[9] J. Deng, K. Li, M. Do, H. Su, and L. Fei-Fei, "Construction and Analysis of a Large Scale Image Ontology," in *Vision Sciences Society*, 2009.

[10] C Fellbaum, *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA, 1998.

[11] F Chollet, "VGG16 model for keras," https://github.com/fchollet/keras/, 2017, Accessed: 2017-09-14.

[12] G Neubig, C Dyer, Y Goldberg, A Matthews, W Ammar, A Anastasopoulos, M Ballesteros, D Chiang, D Clothiaux, T Cohn, K Duh, M Faruqui, C Gan, D Garrette, Y Ji, L Kong, A Kuncoro, G Kumar, C Malaviya, P Michel, Y Oda, M Richardson, N Saphra, S Swayamdipta, and P Yin, "DyNet: The dynamic neural network toolkit," https://arxiv.org/pdf/1701.03980.pdf, 2017, Accessed: 2017-09-14.

[13] P Arthur, G Neubig, and S Nakamura, "Incorporating discrete translation lexicons into neural machine translation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[14] T Fukada, K Tokuda, T Kobayashi, and S Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Internat. Conf. Acoust. Speech and Sign. Process. (ICASSP)*, 1992.

[15] T Yoshimura, K Tokuda, T Masuko, T Kobayashi, and T Kitamura, "Mixed excitation for HMM-based speech synthesis," in *Proc. Eurospeech*, 2001, pp. 2263–2266.

[16] K Tokuda, T Yoshimura, T Masuko, T Kobayashi, and T Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, 2000.

[17] S-H Chen, S-H Hwang, and Y-R Wang, "An RNN-based prosodic information synthesizer for Mandarin text-to-speech," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 3, pp. 226–239, 1998.

[18] AW Black and PK Muthukumar, "Random forests for statistical speech synthesis," in *Proc. Interspeech*, 2015, pp. 1211–1215.

[19] PK Muthukumar and AW Black, "Automatic discovery of a phonetic inventory for unwritten languages for statistical speech synthesis," in *Proc. ICASSP*, 2014.

[20] C Rashtchian, P Young, M Hodosh, and J Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[21] D Harwath and J Glass, "Deep multimodal semantic embeddings for speech and images," in *2015 IEEE Automatic Speech Recognition and Understanding Workshop*, Scottsdale, Arizona, USA, 2015, pp. 237–244.

[22] T-Y Lin, M Maire, S Belongie, J Hays, P Perona, D Ramanan, P Dollár, and CL Zitnick, "Microsoft COCO: Common objects in context," *Lecture Notes in Computer Science*, vol. 8693, pp. 740–755, 2014.

[23] X Chen, H Fang, T-Y Lin, R Vedantam, S Gupta, P Dollar, and C: Zitnick, "Microsoft COCO captions: Data collection and evaluation server," https://arxiv.org/abs/1504.00325, 2015, Downloaded 2017-09-14.

[24] W Havard, L Besacier, and O Rosec, "SPEECH-COCO: 600k visually grounded spoken captions aligned to MSCOCO data set," in *ISCA Workshop on Grounding Language Understanding (GLU2017)*, 2017.

[25] 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 *Internat. Worksh. Spoken Language Translation (IWSLT)*, Lake Tahoe, 2014, pp. 73–79.

[26] O Scharenborg, F Ciannella, S Palaskar, A Black, F Metze, L Ondel, and M Hasegawa-Johnson, "Building an asr system for a low-research language through the adaptation of a high-resource language asr system: Preliminary results," in review, 2017.

[27] L Ondel, L Burget, and J Černocký, "Variational inference for acoustic unit discovery," *Procedia Computer Science*, vol. 2016, no. 81, pp. 80–86, 2016.

[28] K Kurihara, M Welling, and YW Teh, "Collapsed variational dirichlet process mixture models," in *Proc. IJCAI*, 2007.