Speech Technology for Unwritten Languages

Odette Scharenborg, *Senior Member, IEEE*, Laurent Besacier, *Senior Member, IEEE*, Alan Black, *Senior Member, IEEE*, Mark Hasegawa-Johnson, *Senior Member, IEEE*, Florian Metze, *Senior Member, IEEE*, Graham Neubig, Sebastian Stüker, *Member, IEEE*, Pierre Godard, *Member, IEEE*, Markus Müller, *Member, IEEE*, Lucas Ondel, *Member, IEEE*, Shruti Palaskar, *Member, IEEE*, Philip Arthur, *Member, IEEE*, Francesco Ciannella, Mingxing Du, Elin Larsen, Danny Merkx, Rachid Riad, Liming Wang, and Emmanuel Dupoux, *Senior Member, IEEE*

***Abstract*—Speech technology plays an important role in our everyday life. Among others, speech is used for human-computer interaction, for instance for information retrieval and on-line shopping. In the case of an unwritten language, however, speech technology is unfortunately difficult to create, because it cannot be created by the standard combination of pre-trained speech-to-text and text-to-speech subsystems. The research presented in this article takes the first steps towards speech technology for unwritten languages. Specifically, the aim of this work was 1) to learn speech-to-meaning representations without using text as an intermediate representation, and 2) to test the sufficiency of the learned representations to regenerate speech or translated text, or to retrieve images that depict the meaning of an utterance in an unwritten language. The results suggest that building systems that go directly from speech-to-meaning and from meaning-to-speech, bypassing the need for text, is possible.**

***Index Terms*—Speech processing, automatic speech recognition, unsupervised learning, speech synthesis, image retrieval.**

I. Introduction

Speech-enabled devices are all around us, e.g., all smart phones are speech-enabled, as are the smart speakers in our homes. Such devices are crucial when one can only communicate via voice, e.g., when one's eyes and/or hands are busy or disabled, or when one cannot type a query in the native language because the language does not have an orthography or does not use it in a consistent fashion. These languages are typically referred to as unwritten languages. However, for only about 1% of the world languages the minimum amount of transcribed speech training data that is needed to develop automatic speech recognition (ASR) technology is available [1], [41]. Languages lacking such resources are typically referred to as 'low-resource languages', and include, by definition, all unwritten languages. Consequently, millions of people in the world are not able to use speech-enabled devices in their native language. They thus cannot use the same services and applications as persons who speak a language for which such technology is developed, or they are forced to speak in another language.

Much progress in speech-to-text (i.e., ASR) and text-to-speech (i.e., speech synthesis) technology has been driven by the speech-to-text conversion paradigm (e.g., [37]). In this paradigm, all aspects of the speech signal that cannot be converted to text (personality, prosody, performance, emotion, dialect and sociolect, reverberation, environment, etc.) are treated as sources of undesirable variability, and compensated using feature and model normalization methods, for the purpose of focusing energy on a clear and solvable task. To that end, acoustic models of speech sounds are created which are statistical representations of each sound (or phone), in principle devoid of all aspects that cannot be converted to text. Until about 2015,

Manuscript received August 5, 2019; revised December 2, 2019 and January 30, 2020; accepted February 3, 2020. Date of publication February 13, 2020; date of current version February 19, 2020. This work was supported in part by JHU and in part CMU via grants from Google, Microsoft, Amazon, Facebook, and Apple. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which was supported by NSF under Grant OCI-1053575. Specifically, it used the Bridges system, which was supported by NSF Award ACI-1445606, at the Pittsburgh Supercomputing Center (PSC). The work of O. Scharenborg was supported in part by a Vidi-grant from NWO under Grant 276-89-003 and in part by a Delft Technology Fellowship from Delft University of Technology. The work of Pierre Godard, Markus Müller, and Sebastian Stüker was supported in part by the French ANR and in part by the German DFG under Grant ANR-14-CE35-0002 (BULB project). The work of Mingxing Du, Elin Larsen, Rachid Riad, and Emmanuel Dupoux was supported in part by the European Research Council under Grant ERC-2011-AdG-295810 BOOTPHON, and in part by Grants ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL*. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiaodong Cui. *(Corresponding author: Odette Scharenborg.)*

Odette Scharenborg was with the Centre for Language Studies, Radboud University, 6525 XZ Nijmegen, The Netherlands. She is now with the Multimedia Computing Group, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: o.e.scharenborg@tudelft.nl).

Laurent Besacier is with LIG, University Grenoble Alpes (UGA), 38400 Saint-Martin-d'Hères, France (e-mail: laurent.besacier@imag.fr).

Alan Black, Florian Metze, Graham Neubig, Shruti Palaskar, Philip Arthur, and Francesco Ciannella are with Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: awb@cs.cmu.edu; fmetze@cs.cmu.edu; gneubig@cs.cmu.edu; spalaska@andrew.cmu.edu; philip.arthur30@gmail.com; francesco.ciannella@gmail.com).

Mark Hasegawa-Johnson and Liming Wang are with the Beckman Institute, University of Illinois, Urbana-Champaign IL 61801 USA (e-mail: jhasegaw@illinois.edu; lwang114@illinois.edu).

Sebastian Stüker and Markus Müller are with the Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany (e-mail: sebastian.stueker@kit.edu; m.mueller@kit.edu).

Pierre Godard was with LIMSI, 91400 Orsay, France (e-mail: godard@limsi.fr).

Lucas Ondel is with the Brno University, Brno 612 00, Czech Republic, and also with Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: lucas.ondel@gmail.com).

Mingxing Du, Elin Larsen, Rachid Riad, and Emmanuel Dupoux are with ENS/CNRS/EHESS/INRIA, 75005 Paris, France (e-mail: fdsyen@gmail.com; elin_larsen1@hotmail.fr; riadrachid3@gmail.com; emmanuel.dupoux@gmail.com).

Danny Merkx is with the Radboud University, 6525 XZ Nijmegen, The Netherlands (e-mail: d.merkx@let.ru.nl).

Digital Object Identifier 10.1109/TASLP.2020.2973896

the majority of speech-to-text systems required a pronunciation lexicon, and lexicon-based speech-to-text systems still dominate the field. In lexicon-based speech-to-text systems, words (the intended output) typically are modeled as sequences of acoustic models of phones [19], [38], [54]. In text-to-speech systems, the lexicon determines the order of context-dependent phone models, and the context-dependent phone models specify the process by which the acoustic signal is generated [67], [68]. Recent end-to-end deep neural networks usually bypass phones, in order to convert audio input directly into text output [10], [22], [62]. Both phone-based and end-to-end systems, however, for both speech-to-text and text-to-speech conversion, require text: it is necessary to train the statistical model and/or neural network using a large (sometimes very large) database of audio files with matched text transcriptions.

In the case of an unwritten language, text cannot be used, and the speech-to-text and text-to-speech technology thus needs to be modified. Methods for doing so may be guided by early work on speech understanding, when text was considered to be a stepping stone on the path between speech and meaning [47]. Training a speech-to-meaning system is difficult, because few training corpora exist that include utterances matched to explicit semantic parse structures; the experiences reported in [32] suggest that such corpora are expensive to create. On the other hand, a semantic parse is not the only way to communicate the meaning of an utterance.

Ogden and Richards [51] defined meaning to be a three-part relationship between a reference (a "thought" or cognitive construct), a referent (a physical object which is an "adequate" referent of the reference), and a symbol (a physical sign which is defined, in some linguistic system, to be "true" if and only if it connects to an existing cognitive reference in the mind of the speaker or writer). In their model, the reference is never physically observable, because it exists only in the mind of the speaker. Communication between two humans takes place by the use of symbols (speech signals or written symbols), possibly with the help of gestures pointing to adequate referents (physical objects or pictures). Consider the model of semantics shown in Fig. 1. In this model, the reference (the logical propositional form of an utterance's meaning) is unknown, but instead, we have two different symbols (a spoken utterance in one language, and a text translation in another language) and one referent (an image considered by at least one transcriber to be an adequate depiction), all linked to the same reference. Suppose we have a corpus in which some utterances are matched to translations in another (written) language, some to images, and some to both; can we learn a representation of the meaning of the sentence that is sufficient to regenerate speech, a translation, and/or retrieve an image from a database?

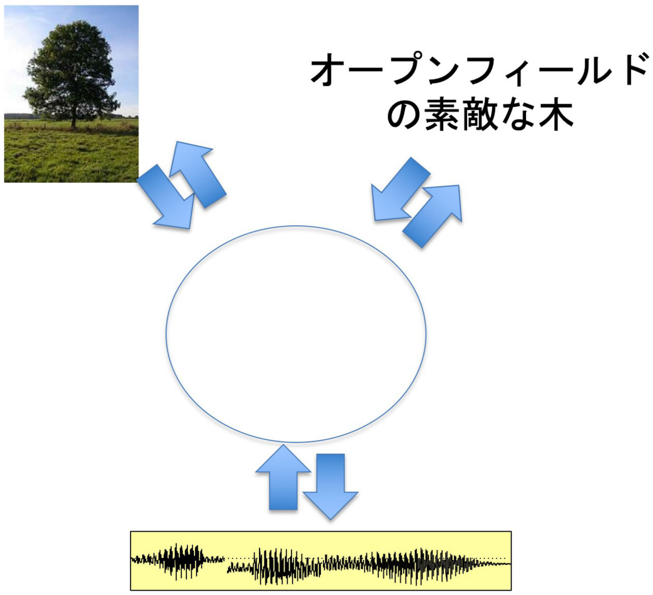


Fig. 1. A model of semantics for speech technology development in an unwritten language. The speech signal (bottom of the figure) has some propositional content which is unknown and not directly observable (represented by the open circle in the center of the figure). Instead of directly observing the propositional meaning of the utterance, it is possible to observe its translation to another language (top right, i.e., in Japanese), or to observe an image depicting the meaning of the utterance (top left).

To answer this question, we present three speech technology applications that might be useful in an unwritten language situation. The first task is end-to-end (E2E) speech-to-translation. In this task, a translation is created from raw speech of an unwritten language into a textual transcription of another language without any intermediate transcription [5], [72]. This technology is attractive for language documentation, where corpora are created and used consisting of audio recordings in the language being documented (the unwritten, source language) aligned with their translations in another (written) language, without a transcript in the source language [1], [7]. The second task is speech-to-image retrieval. Speech-to-image retrieval is a relatively new task [2], [23], [27], in which images and speech are mapped to the same embedding space, and an image is retrieved from an image database using spoken captions. While doing so, the system uses multi-modal input to discover speech units in an unsupervised manner. In a way, this is similar to how children acquire their first language. Children learn a first language using different modalities including the visual modality and the auditory modality. Learning can then occur in both a supervised way (e.g., a caretaker saying "There is a ball" while simultaneously and explicitly showing a ball to the child) and in an unsupervised manner (i.e., without explicit referents or explicitly turning the child's attention to an object, e.g., by talking about a ball without pointing at it). This technology is attractive for, e.g., online shopping. A user might be interested in buying a coat, and ask for images of coats. The third task is image-to-speech. Image-to-speech is a new speech technology task [28], [29], which is similar to automatic image captioning, but can reach people whose language does not have a natural or easily used written form. An image-to-speech system should generate a spoken description of an image directly, without first generating text. This technology could be interesting for social media applications. Particularly in situations where the receiver of an image is not able to look at a screen, e.g., while driving a car. The speech-to-image and speech-to-translation tasks bypass the need for traditional phone-based acoustic models trained on large databases of speech, and instead map the speech directly to the image or translation. The image-to-speech application creates "acoustic models" by automatically discovering speech units from the speech stream. All three systems use a common encoder-decoder architecture, in which the sequence of inputs is encoded to a latent semantic space, permuted along the time axis using a neural attention mechanism, then decoded into a different modality. Note that all three experiments use similar methods to compute the semantic encoding, but the encoding weights are separately optimized for each application. The possibility of sharing a single encoding between experiments was not explored in this paper, but could be the subject of future research.

The remainder of this paper describes the systems that learn an underlying semantic representation in order to regenerate the speech signal, or its text translation, or to retrieve an image that depicts the same propositional content from a database. Section II describes relevant background. Section III describes the Deep Neural Network (DNN) architectures used for all experimental and baseline systems. Section IV describes the databases used for the experiments, and the methods used to train and test the speech-to-translation, speech-to-image, and image-to-speech systems. Section V gives experimental results, Section VI is discussion, and Section VII concludes.

II. Background

Algorithms for speech-to-translation generation, image-to-speech generation, and speech-to-image retrieval have previously been published separately by a number of different authors. To the best of our knowledge, this is the first paper seeking to develop a unified framework for the generation of all three types of speech technology for unwritten languages.[1]

Speech-to-translation for unwritten languages was first proposed in [6]; E2E neural machine translation methods for this task were first described in [5], [16]. The 2018 International Workshop on Spoken Language Translation (IWSLT) was the first international competition that evaluated systems based on E2E speech-to-text translation performance, without separately evaluating text transcription in the source language [50]. Most participants in the IWSLT competition still relied on separately trained speech recognition and machine translation subsystems ("pipelined systems"), but at least two papers described neural machine translation systems trained E2E from speech in the source language to text in the target language [15], [35]. The E2E systems were however outperformed by the pipelined system: [35] reported BLEU scores of 14.87 for the pipelined system, and of 4.44 for the E2E system; although transfer learning from the pipelined to the E2E system improved its BLEU from 4.44 to 6.71. The transfer learning idea was further developed in [4] by first training a speech recognizer in a written language (English or French), then transferring the parameters of the trained speech encoder to the input side of a speech-to-translation system for an unwritten language (Spanish or Mboshi). Significant improvements (of 11.60 BLEU) were also obtained by fine-tuning the E2E system using cleaned subsets of the training data [15].

The image-to-speech generation task was proposed in [28], [29], and consists of the automatic generation of a spoken description of an input image. The methods are similar to those of image captioning, but with speech instead of text outputs. Image captioning was first defined to be the task of generating keywords to match an image [55]. The task of generating keywords from an image led to alternate definitions using text summarization techniques [58] and image-to-text retrieval techniques [33]. End-to-end neural image captioning (using text), using an output LSTM whose context vectors are attention-weighted summaries of convolutional inputs, was first proposed in [75].

While the speech-to-translation and image-to-speech tasks described in this paper are both generation tasks: the output (text or speech, respectively) is generated by a neural network, to our knowledge, no similar generation network has yet been proposed for the speech-to-image task. Instead, experiments in this paper are based on the speech-to-image retrieval paradigm, in which spoken input is used to search for an image in a predefined large image database [23]. During training, the speech-to-image system is presented with (image,speech) pairs, where the speech signal consisted of spoken descriptions of the image. The speech and images are then projected into the same "semantic" space. The DNN then learns to associate portions of the speech signal with the corresponding regions in the image. For instance, take a stretch of speech containing the words "A nice tree in an open field" (please note, in this paradigm there are no transcriptions available but for ease of reading the acoustic signal is written out in words here, see Fig. 1) and an image of a tree in a grassy field. If the sound of the word "tree" is associated with similar visible objects in a large enough number of training images, the DNN then learns to associate the portion of the acoustic signal which corresponds to "tree" with the region in the image that contains the "tree," and as such is able to learn word-like units and use these learned units to retrieve the image during testing (i.e., image retrieval) [27]. The semantic embedding of input sentences can be further improved by acquiring tri-modal training data, in which each image is paired with a spoken description in one language and a text description in another language; the retrieval system is then trained to compute a sentence embedding that is invariant across the three modalities [24].

III. Architecture

We assume that all three modalities (speech, translated text, and images) can be projected into a common semantic embedding space using convolutional and recurrent encoder networks, and can then be regenerated from the semantic space using decoder networks. We assume that text input is presented in the form of a one-hot embedding. Speech is presented as a sequence of mel-frequency cepstral coefficient (MFCC) vectors. Images are pre-encoded using a very deep convolutional neural network,

[1] Note, a summary and initial results of this work were presented in [59] also available in the HAL repository: [Online]. Available: https://hal.archives-ouvertes.fr/hal-01709578/document. The current paper provides more details on the experimental setups of the experiments, including more details on the used Deep Neural Network architectures and algorithms and rationales for the experiments. Moreover, new results are presented for the image2speech task and the speech2image task for which we report the currently best results compared to results reported in the literature. Additionally, new baseline results are added for the image2speech task compared to [28], [29].

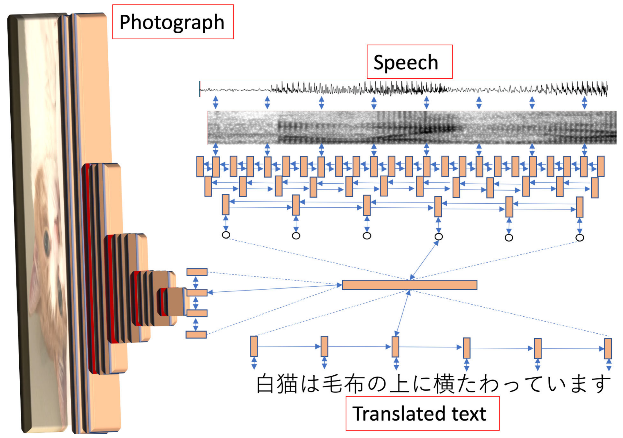


Fig. 2. Proposed neural architecture. Separate encoder and decoder networks are trained for each of the three modalities. The figure shows the speech encoder (a pyramidal LSTM), and decoders (LSTMs with attention-weighted input context vectors) that would generate an image output or a translated text ouput.

with weights pre-trained for the ImageNet image classification task, e.g., using the VGG16 [63] implementation of [20]. In order to convert the image into a sequence of vectors appropriate for encoding by a recurrent neural network, the penultimate feature map of the ImageNet classifier is converted into a two-dimensional array of sub-images (overlapping regions of 40×40 pixels each), which is then read in raster-scan order, one row after another, in order to form a one-dimensional pseudo-temporal sequence. These assumptions are satisfied by the architecture shown in Fig. 2.

Let $X = [\vec{x}_1, \ldots, \vec{x}_{T_X}]$ be a sequence of T_X MFCC vectors representing the speech utterance, let $Y = [\vec{y}_1, \ldots, \vec{y}_{T_Y}]$ be a sequence of T_Y one-hot vectors representing the translated text, and let $Z = [\vec{z}_1, \ldots, \vec{z}_{T_Z}]$ be a sequence of feature vectors representing overlapping sub-images in raster-scan order. The problem of speech-to-translation generation, then, is to learn a function f_{YX} that minimizes a loss function $\mathcal{L}(Y, f_{YX}(X))$. The problem of image-to-speech generation is to learn a function f_{XZ} that minimizes a similar loss function, $\mathcal{L}(X, f_{XZ}(Z))$. The problem of speech-to-image retrieval (or image-to-speech retrieval) is to learn embedding functions $g_X(X)$ and $g_Z(Z)$ in order to minimize a pair-wise loss function between correct retrieval results, $\mathcal{L}(g_X(X), g_Z(Z))$.

The architecture shown in Fig. 2 represents the speech-to-translation task, f_{YX}, as the composition of a speech encoder g_X and a text decoder h_Y. Likewise, the image-to-speech task, f_{XZ}, is the composition of g_Z and h_X, thus

$$f_{YX}(X) = h_Y(g_X(X)), \quad \text{and} \quad f_{XZ}(Z) = h_X(g_Z(Z)) \quad (1)$$

The speech encoder, g_X, is modeled as a pyramidal bidirectional long-short term memory network (pyramidal biLSTM): a biLSTM with three hidden layers, in which the input to each layer is the concatenation of two consecutive state vectors from the layer below (thus each layer has half as many frames as the layer below it). The speech-to-image retrieval system uses as its image encoder, g_Z, a pre-trained deep convolutional neural net. The image-to-speech generation system uses the same pre-trained convolutional network, followed by a three-layer pyramidal biLSTM. For image coding purposes, the input to the biLSTM is created by scanning the last convolutional layer in raster-scan order, i.e., left-to-right, top-to-bottom. For example, in the VGG16 encoder [20], [63] used in our image-to-speech experiments and our initial speech-to-image experiments, the last convolutional layer has 512 channels, each of which is 14×14, therefore $\vec{z}_t$ is a 512-dimensional vector, and $T_Z = 14 \times 14 = 196$.

For purposes of more detailed exposition, consider the image-to-speech system, $f_{XZ}(Z) = h_X(g_Z(Z))$. The encoder is a pyramidal biLSTM with three hidden layers, and with each layer downsampled by a factor of two relative to the preceding layer. For example, $\vec{e}_{l,t}$, the t^{th} LSTM state vector at level l of the network, is computed from the preceding time step ($\vec{e}_{l,t-1}$) and the preceding level ($\vec{e}_{l-1,2t-1}$ and $\vec{e}_{l-1,2t}$):

$$\vec{e}_{l,t} = \gamma\left(\vec{e}_{l,t-1}, \vec{e}_{l-1,2t-1}, \vec{e}_{l-1,2t}\right) \quad (2)$$

The output is a sequence of encoder state vectors at the L^{th} level,

$$g_Z(Z) = \left[\vec{e}_{L,1}, \ldots, \vec{e}_{L,D_Z}\right], \quad (3)$$

where $D_Z = T_Z 2^{-L}$ is the number of state vectors in the L^{th} level of the encoder.

The speech decoder, h_X, has two parts. In the first part of the decoder, the embedding sequence $g_Z(Z)$ is converted into a sequence of monophone labels by an LSTM. In the second part of the decoder, the monophone sequence is converted into a sequence of MFCC vectors (for more details, please see Section IV.C.4), $X = [\vec{x}_1, \ldots, \vec{x}_{T_X}]$, by a random forest regression algorithm. The first part of the decoder is an LSTM, whose inputs are attention-weighted context vectors, $\vec{c}_i$, computed from the encoder state vectors as

$$\vec{c}_i = \sum_{t=1}^{D_Z} a_{it} \vec{e}_{L,t}, \quad (4)$$

where a_{it} is the attention weight connecting the i^{th} decoder state vector, $\vec{s}_i$, to the t^{th} encoder time-step, $\vec{e}_{L,t}$, and is computed by a two-layer feedforward neural net $\alpha(\vec{s}_{i-1}, \vec{e}_{L,t})$ as

$$a_{it} = \frac{\exp \alpha(\vec{s}_{i-1}, \vec{e}_{L,t})}{\sum_{\tau=1}^{D_Z} \exp \alpha(\vec{s}_{i-1}, \vec{e}_{L,\tau})}. \quad (5)$$

The decoder state vectors are generated by a single LSTM layer, β, as

$$\vec{s}_i = \beta\left(\vec{s}_{i-1}, \vec{c}_i, \hat{y}_{i-1}\right) \quad (6)$$

The probability of the monophone j being computed as the i^{th} output symbol, $\Pr(m_i = j)$, is computed using a softmax layer, in which the LSTM state vector $\vec{s}_i$ and context vector, $\vec{c}_i$, are concatenated, multiplied by a weight vector $\vec{w}_j$, and normalized so that the output is a probability mass function:

$$\Pr(m_i = j | m_1, \ldots, m_{i-1}, Z) = \frac{\exp([\vec{s}_i^T, \vec{c}_i^T]\vec{w}_j)}{\sum_k \exp([\vec{s}_i^T, \vec{c}_i^T]\vec{w}_k)} \quad (7)$$

Since the state vector $\vec{s}_i$ is a function of all preceding output symbols $[\hat{y}_1, \ldots, \hat{y}_{i-1}]$, it is possible that a high-probability output in any given frame might lead to low-probability outputs in future frames; to ameliorate this problem, we used a Viterbi beam search with a beamwidth of 20. The resulting monophone sequence is used as input to a random forest in order to compute both the duration of each monphone, and the sequence of MFCC vectors $X = [\vec{x}_1, \ldots, \vec{x}_{T_X}]$.

The image-to-speech neural network components are trained to minimize the cross-entropy between the generated monophone sequence and the reference monophone sequence. The random forest is trained, using the FestVox algorithm [8], to minimize mean cepstral distortion between the reference MFCC and the MFCC generated by the random forest from the reference monophone sequence. The speech-to-translation neural network is trained to minimize cross-entropy between the softmax output probabilities $\Pr(\hat{y}_i = j)$ (as in Eq. (7), but computing the probability of an output word $\hat{y}_i$, rather than the probability of an output monophone m_i) and the reference translated word sequence $Y = [\vec{y}_1, \ldots, \vec{y}_{T_Y}]$:

$$\mathcal{L}\left(Y, f_{YX}(X)\right) = -\sum_{i=1}^{T_Y} \ln \Pr(\hat{y}_i = \vec{y}_i | \hat{y}_1, \ldots, \hat{y}_{i-1}, X) \quad (8)$$

The speech-to-image retrieval task requires us to measure the similarity between the embedding of any particular speech sequence, $g_X(X)$, and the corresponding image sequence, Z. Speech-to-image retrieval experiments were tested using LSTM encoders for both speech and image, but in the end, the best-performing system used a fully-connected image encoder, which we will denote $\gamma_Z(Z)$, rather than the LSTM image encoder $g_Z(Z)$ described in Eq. (3). The fully-connected LSTM encoder first performs 2×2 max-pooling in each of the 512 channels of the last convolutional layer, in order to create a tensor of size $7 \times 7 \times 512$; this tensor is then flattened into a vector of length $7 \times 7 \times 512 = 25088$, and transformed through three fully-connected layers to create a vector $\gamma_Z(Z)$ with 1024 dimensions. In some experiments, the image embedding $\gamma_Z(Z)$ was computed from the same VGG-based CNN features as the image-to-speech system; in the most successful experiments, it was computed, instead, from a different pre-trained CNN (Resnet-152 [31]). In both cases, the speech encoder is a bidirectional recurrent neural network (RNN) with architecture similar to that described in Eq. (3); early experiments used the same three-layer pyramidal biLSTM as the speech-to-translation system, but the most successful experiments used, instead, a network with one convolutional layer followed by a bidirectional GRU. In either case, the state vectors of the speech RNN, $\vec{e}_{L,t}$, were combined using attention weights, a_{Zt}, computed as a measure of the similarity between the speech state vector and the fixed-length image embedding vector $\gamma_Z(Z)$, in order to create a context vector $\vec{c}_{ZX}$:

$$\vec{c}_{ZX} = \sum_{t=1}^{T_X} a_{Zt} \vec{e}_{L,t} \quad (9)$$

$$a_{Zt} = \frac{\exp \alpha(\gamma_Z(Z), \vec{e}_{L,t})}{\sum_{\tau=1}^{D_X} \exp \alpha(\gamma_Z(Z), \vec{e}_{L,\tau})} \quad (10)$$

where $\alpha()$ is a two-layer fully-connected feedforward network with the same architecture as the $\alpha()$ network in Eq. (5). For any particular speech signal, the speech-to-image system returns the image that maximizes the cosine similarity measure $\cos(X, Z)$, defined to be

$$\cos\left(X, Z\right) = \frac{\vec{c}_{ZX}^T \gamma_Z(Z)}{\|\vec{c}_{ZX}\| \cdot \|\gamma_Z(Z)\|} \quad (11)$$

The network weights are then trained using a bi-modal triplet loss. The bi-modal triplet loss was defined by [23] to be similar to a standard triplet loss [11], but with incorrect exemplars $X' \neq X$ and $Z' \neq Z$ drawn uniformly at random for both the speech and image modalities. The loss is then computed as the sum, over all correct pairs (X, Z) in the minibatch B, of the clipped difference between similarities of the incorrect and correct pairs:

$$\mathcal{L} = \sum_{(X,Z),(X,'Z')\in B} \Big(\max(0, \cos\left(X, Z'\right) - \cos\left(X, Z\right) + 1) + \max(0, \cos\left(Z, X'\right) - \cos\left(Z, X\right) + 1) \Big) \quad (12)$$

IV. Experimental Set-up

Fig. 1 suggests a three-part model of semantics, in which the meaning of an utterance (its cognitive representation) is unknown, but is indicated by a text translation and by an image referent. In order to test the model, it is necessary to acquire training and test data, and to define training and test evaluation criteria.

A complete test of Fig. 1 requires data in which each utterance is matched to a text translation, and to an image. Such data exist in no unwritten language, therefore some type of proxy dataset is necessary. Two types of proxy datasets are used in this paper: a proxy dataset containing all three modalities, but with speech in a language that is not truly unwritten (FlickR-real), and a proxy dataset containing only two modalities (speech and translation), with speech in a language that is truly unwritten (Mboshi).

First, the FlickR-real speech database is a tri-modal (speech, translated text, images) corpus, but the speech is in a language that is not truly unwritten nor a low-resource language (English). The images in this dataset were selected through user queries for specific objects and actions from the FlickR photo sharing website [33]. Each image contains five descriptions in natural language which were collected using a crowdsourcing platform (Amazon Mechanical Turk; AMT) [33]. AMT was also used by [23] to obtain 40 K spoken versions of the captions. These are made available online.[2] We augmented this corpus in two ways. First, the database was made tri-modal by adding Japanese translations (Google MT [73]) for all 40 K captions, as well as Japanese tokenization. Second, we generated monophone transcriptions of all English speech files: original text prompts were converted to monophone sequences using CMUdict [42], after which the original text prompts were discarded, and not used for any further purpose. Other than the monophone transcriptions, no other English-language resources were used; thus, apart from the monophone transcriptions, English was treated as an unwritten language.

[2] [Online]. Available: https://groups.csail.mit.edu/sls/downloads/flickraudio/

TABLE I
OVERVIEW OF THE DATABASES

Data set	Language	Size	Paired translations	Paired images	#spkrs	Tasks
Mboshi	Mboshi	5h	yes (French - Human)	no	3	Speech-to-translation
FlickR-real speech	English	62h	yes (Japanese - MT)	yes	183	All three tasks

The second proxy dataset used a truly unwritten language (Mboshi), but contained only two of our target modalities (speech and translation). Mboshi is a Bantu language (Bantu C25) of Congo-Brazzaville [1], [64]. Mboshi was chosen as a test language because Mboshi utterances and their paired French translations were available to us through the BULB project [1]. The Mboshi corpus [21] was collected using a real language documentation scenario, using ligaikuma,[3] a recording application for language documentation [7]. The Mboshi corpus is a multilingual corpus consisting of 5 K speech utterances (approximately 4 hours of speech) in Mboshi with hand-checked French text translations. Additionally, the corpus contains linguists' monophone transcriptions (in a non-standard graphemic form which was designed, by the linguists who used it, to represent the phonology of the language) [1], [21]. The corpus is augmented with automatic forced-alignments between the Mboshi speech and the linguists' monophone transcriptions [13]. The corpus and forced alignments are made available to the research community.[4] Monophone transcriptions of the Mboshi corpus were used in order to train and test translations from Mboshi speech to Mboshi monophone sequences, but were not used for the translation of Mboshi speech to French text (see below).

Table I gives an overview of the characteristics of the multimodal datasets, which were used in the experiments.

The neural architecture shown in Fig. 2 was trained using the XNMT [17], [48] architecture, and tested in three applications: speech-to-translation sequence generation, speech-to-image retrieval, and image-to-speech sequence generation.

A. Speech-to-Translation

We built end-to-end speech-to-translation systems with the neural sequence-to-sequence machine translation toolkit XNMT [17], [48] on the FlickR-real (English-to-Japanese) and Mboshi corpora (Mboshi-to-French). The speech-to-translation systems were based on the neural machine translation functionality [3], [39], [49], [65] of XNMT.

The speech encoder for the speech-to-translation experiments (Fig. 3) takes in a sequence of speech feature vectors, and converts them into a format conducive for translation. The encoder used a bi-directional pyramidal LSTM. The first layer observes speech features computed by a convolutional neural network applied over MFCCs inputs.

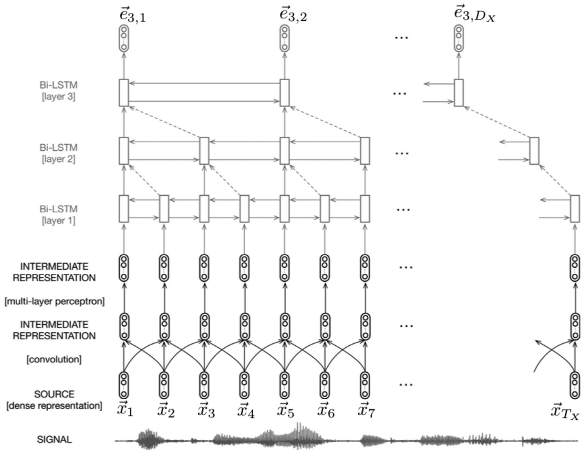

$\vec{e}_{3,1}$
$\vec{e}_{3,2}$
...
$\vec{e}_{3,D_X}$
Bi-LSTM [layer 3]
Bi-LSTM [layer 2]
Bi-LSTM [layer 1]
INTERMEDIATE REPRESENTATION
[multi-layer perceptron]
INTERMEDIATE REPRESENTATION
[convolution]
SOURCE [dense representation]
$\vec{x}_1$ $\vec{x}_2$ $\vec{x}_3$ $\vec{x}_4$ $\vec{x}_5$ $\vec{x}_6$ $\vec{x}_7$
$\vec{x}_{T_X}$
SIGNAL

Fig. 3. The encoder architecture for the speech-to-translation experiments was a three-layer bi-directional pyramidal LSTM, observing speech features computed by a one-layer convolutional network over the top of MFCCs.

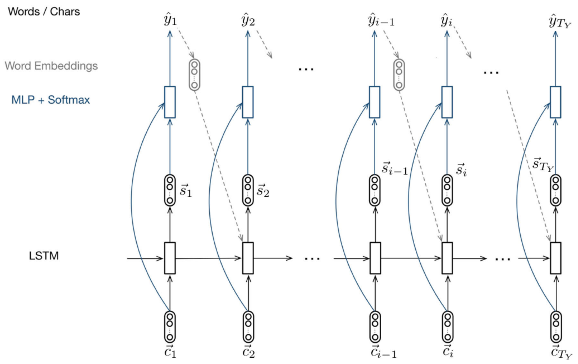

Words / Chars
Word Embeddings
MLP + Softmax
LSTM
$\hat{y}_1$ $\hat{y}_2$ $\hat{y}_{i-1}$ $\hat{y}_i$ $\hat{y}_{T_Y}$
...
$\vec{s}_1$ $\vec{s}_2$ $\vec{s}_{i-1}$ $\vec{s}_i$ $\vec{s}_{T_Y}$
$\vec{c}_1$ $\vec{c}_2$ $\vec{c}_{i-1}$ $\vec{c}_i$ $\vec{c}_{T_Y}$

Fig. 4. The decoder architecture for speech-to-translation experiments was a one-layer LSTM generating characters as output (word outputs were also tested, but were not as successful).

The decoder, shown in Fig. 4, is an LSTM that generates either word or character outputs. Word-output systems always exhibited lower BLEU scores (both word-based BLEU and character-based BLEU), therefore results will only be reported for systems that generated character outputs. The decoder is a uni-directional LSTM, observing context vectors c_i that are generated by the attention-weighted combination of input encoder vectors. Each LSTM cell also observed the previous frame's LSTM cell, and a one-hot vector specifying the identity of the character generated in the previous frame.

The encoder and decoder are combined to generate an output sentence character-by-character in a probabilistic fashion, given the spoken input sentence. During training, the model's parameters are updated using stochastic gradient descent on the cross-entropy loss computed from the training corpus; training stops when cross-entropy of an independent validation set stops decreasing.

[3] [Online]. Available: http://lig-aikuma.imag.fr

[4] It is made available for free from ELRA at: [Online]. Available: http://catalogue.elra.info/en-us/repository/browse/ELRA-S0396/; it can also be retrieved online at: [Online]. Available: https://github.com/besacier/mboshi-french-parallel-corpus

B. Speech-to-Image

The speech-to-image retrieval system was implemented in PyTorch. Image referents are not available for the Mboshi corpus, therefore speech-to-image experiments were only performed using FlickR-real. The training set consisted of 6000 training images, 1000 test, and 1000 validation images. When an image is part of the training or validation corpus, all of its spoken captions are used, thus the FlickR-real training corpus included 30,000 audio-image pairs (6000 distinct images).

The model used a pretrained ResNet-152 [31] with the top layer removed to encode the images. These features were then fed into a fully connected layer with 1024 units. The speech was encoded using a 1 d convolutional layer with stride 2, width 6 and 64 output channels on the MFCCs. The resulting features were fed into a GRU with 1024 hidden units and finally a vectorial self-attention layer [12]. The resulting embeddings were normalized to have unit L2 norm, and used a similarity score based on cosine similarity (Eq. (11)) between the image and speech embeddings to perform the retrieval task.

Two types of acoustic features were compared: 1) MFCCs (baseline features), similar to [23] but with added speaker-dependent mean-variance normalization on the features before zero-padding/truncation. We used 10 ms skip step and 25 ms window for the spectrogram and 40 filters; 2) Multilingual Bottleneck features (MBN) [18]. The MBN were taken from the hidden layer of a neural network trained on multiple source languages in order to learn a multilingual feature space more generally applicable to all languages. The MBN features are extracted at a 10 ms rate from a 80-dimensional bottleneck layer of a feed-forward neural network trained to classify the senones of multiple languages. The neural network was trained on 17 languages [18]; none of them English. Although the MBN feature is supervised, it does not require any text transcription of the target language.

C. Image-to-Speech

The image-to-speech pipeline [29] consists of four types of standard open-source software toolkits: 1) an **image embedder** VGG16 visual object recognizer which converts each image into a sequence of image feature vectors, 2) a **speech segmenter** that discovers discrete phone-like speech segments in the unwritten language, 3) an **image-to-segment** transducer that learns, and then implements, the mapping from image feature vectors into speech segment labels, and 4) a **segment-to-speech** transducer that learns, and then implements, the mapping from speech segment labels into speech signals. Training data (for the image-to-segment and segment-to-speech transducers) and testing data (for all four components) were drawn from FlickR-real. Each image has five associated speech files, and their associated segment transcriptions. The image-to-segment transducer was trained in order to minimize its average loss, averaged across all five of the segment transcriptions for each training image. The segment-to-speech transducer was separately trained to replicate the segment-to-speech mappings for all training pairs. The FlickR-real training corpus included 6000 images, associated with 30,000 speech files. The validation set consisted of 1000 validation images, associated with 4000 speech files, while a further 1000 images (4000 speech files) were used for testing.

1) Image Embedder: The image embedder was implemented using part of a pre-trained VGG16 object recognizer: the TensorFlow re-implementation, by [20], of the best single network solution [63] in the Imagenet Large Scale Visual Recognition Challenge 2014 Sub-task 2a, “Classification+localization with provided training data,”, which is a 13-layer convolutional neural network trained using the 14 million images of ImageNet [14].

2) Speech Segmenter: Speech segments are monophones, or monophone-like units. Two different systems were tested. First, English-language monophone transcriptions of the FlickR-real corpus were generated from the distributed text prompts (the text prompts were then discarded, and not used for any other purpose). Since English is the language of both the audio and the phone transcripts, these phone transcriptions could be called **same-language** phone transcripts.

Second, in the **cross-language** definition of units approach [60], [61], a DNN was trained on a high-resource language, Dutch, which was subsequently mapped to English (of the FlickR-real database). Although Dutch and English are both Germanic languages, their phoneme inventories differ. The number of Dutch phonemes in the Spoken Dutch corpus (Corpus Gesproken Nederlands, CGN,[5] [53]) that was used for this task is 42, while the number of English phonemes in the FlickR dataset was 45. Eleven Dutch phonemes are not present in English and their corresponding vectors were removed from the soft-max layer. Fifteen English phonemes do not exist in Dutch. Nine of these are diphthongs or affricates which can each be constructed from a sequence of two Dutch phonemes. Six English phonemes, however, need to be created which is done through a linear extrapolation between two (or three) vectors in the soft-max layer corresponding to two (or three depending on the English phoneme which needs to be created) existing Dutch acoustic units (D; see for the mapping [60]). The Dutch vectors that are used to initialize the new English acoustic feature vectors are chosen manually on the basis of their linguistic similarity to the English phonemes which need to be created, e.g., to create English /ae/, an initial vector is created by extrapolating between Dutch /a/ and Dutch /ε/ using:

$$\vec{V}_{\phi,E} = \vec{V}_{\phi,D1} + \alpha(\vec{V}_{\phi,D2} - \vec{V}_{\phi,D3}), \tag{13}$$

where $\vec{V}_{\phi,E}$ is the vector of the missing English phone ϕ, E that needs to be created, $\vec{V}_{\phi,Dx}$ are the vectors of the Dutch phones ϕ, Dx in the soft-max layer that are used to create the vector for the missing English phone ϕ, E. Among the three Dutch phones, D1 refers to the phone which is used as the starting point from which to extrapolate the missing English phone, and D2 and D3 refer to the Dutch phones whose displacement is used as an approximation of the displacement between the Dutch starting

[5] The CGN is a corpus of almost 9 M words of Dutch spoken in the Netherlands and in Flanders (Belgium) in over 14 different speech styles, ranging from formal to informal. For the experiments reported here, we only used the read speech material from the Netherlands, which amounts to 551,624 words for a total duration of approximately 64 hours of speech.

vector and the English vector that should be created. α is a factor corresponding to the approximation of the displacement of $\vec{V}_{\phi,E}$ from $\vec{V}_{\phi,D1}$ and was set manually.

Subsequently, the DNN trained on Dutch but with output vectors adapted to English is used to decode the FlickR-real English data, creating so-called 'self-labels'. The thus-obtained self-labels of the English data are used to retrain the Dutch DNN towards English. The DNN is then iteratively retrained with the English self-labels.

3) Image-to-Segment Transducer: The mapping from image features to segment sequences (same-language phones or cross-language phones) is learned, and then implemented during test time, using a sequence-to-sequence neural network implemented in XNMT. The image-to-segment model learned by XNMT is a sequence-to-sequence model, composed of an encoder, an attender, and a decoder. The encoder has three 128-dimensional bidirectional LSTM hidden layers: the input of each layer is the concatenation of two sequential outputs from the previous layer, so that the time scale decreases by a factor of two with each layer. The input to the encoder are the image feature vectors created by the VGG16 object recognizer. The attender is a two-layer perceptron. For each combination of an input LSTM state vector and an output LSTM state vector (128 dimensions each), the attender uses a two-layer perceptron (one hidden layer of 128 nodes) to compute a similarity score. The decoder has one hidden layer, which is a 128-dimensional unidirectional LSTM. The output layer of the decoder is a softmax where each output node is a speech unit in the unwritten language. The number of output nodes is equal to the size of the speech unit vocabulary in the unwritten language. The output of XNMT is a sequence of discrete speech units, e.g., monophones.

4) Segment-to-Speech-Transducer: The mapping from segment sequences to speech signals is learned, and then implemented during test time, using a random forest regression algorithm implemented in Clustergen [8]. The Clustergen speech synthesis algorithm 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. A machine learning algorithm (e.g., regression tree [8] or random forest [9]) is applied to learn the mapping. The mapping is refined, during training, by resynthesizing each speech signal from the learned units, and then aligning the synthetic and original speech waveforms [45]. Clustergen works well with small corpora because it treats each frame of the training corpus as a training example. It is able to generate intelligible synthetic voices from these small training corpora using an arbitrary discrete labeling of the corpus that need not include any traditional type of phoneme [46], which makes it suitable for our low-resource scenario.

V. Results

A. Speech-to-Translation

Four speech-to-translation systems were trained, two same-language and two cross-language systems using two different input languages: English (using audio from the FlickR-real corpus), and Mboshi (using audio from the Mboshi corpus). For each spoken language, two different text outputs were computed: text output in the same language (English or Mboshi), and text output in a different language (English to Japanese, Mboshi to French). Resulting character BLEU scores (average recall accuracy of character 1-gram through 5-gram sequences [56]) are shown in Table II. Word-level BLEU scores were not calculated, because they are essentially zero: there are very few complete and correct words in the generated output. Note, other papers have also reported very low BLEU scores for this task; the highest reported word-level BLEU score for the Mboshi-to-French corpus, of which we are aware, is only 7.1% [4].

TABLE II
Speech-to-Translation Results (Character BLEU Score, %) for the FlickR-Real and Mboshi Corpora. Val = Validation Set, Test = Evaluation Test Set

Speech	Translation	BLEU (%: Val)	BLEU (%: Test)
English	English	17.74	12.71
English	Japanese	30.99	25.36
Mboshi	Mboshi	56.91	39.53
Mboshi	French	22.36	12.28

As Table II shows, the character BLEU scores for English-to-Japanese were significantly higher than those for Mboshi-to-French. Interestingly, the BLEU scores for the same language English-English task were lower than those for the English-Japanese translation task.

B. Speech-to-Image

TABLE III
Speech-to-Image Retrieval Results (Recall@N in %) for the Tested Input Speech Features

Feature type	R@1	R@5	R@10
Alishani et al. [2]	5.5	16.3	25.3
MFCC	7.3	21.8	32.1
Multiling. Bottleneck	**7.6**	**23.9**	**36.0**

Table III shows the results for the two features for the speech-to-image task evaluated in terms of Recall@N. For reference, the best results in the literature to date on the same data set, i.e., those by Alishani and colleagues [2], are added to Table III. As the results clearly show, both the MBN and MFCC based models show state-of-the-art results. The MBN features are superior to the MFCC features, with an improvement of 1.9% absolute for R@1 which increased to 10.7% absolute for R@10 on the previous best results by [2].

C. Image-to-Speech

The image-to-speech system was trained using either same-language phone transcriptions (generated from the English-language prompts distributed with FlickR-real) or cross-language phone transcriptioning (generated by a Dutch ASR, mapped to English phones using a knowledge-based cross-language mapping). The Phone Error Rate (PER) of the cross-language recognizer prior to retraining was 72.59%, which is comparable to the phone error rates (PER) of cross-language

TABLE IV
IMAGE-TO-SPEECH RESULTS (PHONE-LEVEL BLEU SCORES AND PHONE ERROR RATES (PER (%)) ON THE VAL(IDATION) AND TEST SETS OF THE SAME-LANGUAGE AND CROSS-LANGUAGE IMAGE-TO-SEGMENT TRANSDUCERS

System	Val BLEU	Val PER	Test BLEU	Test PER
Human Transcr.	48.4	66.4	48.6	66.1
Same-language	30.7	70.4	30.2	70.4
Cross-language	25.9	71.9	25.9	71.7
Chance	2.9	81.8	3.4	81.5

ASR systems (e.g., [30] reports PER ranging from 59.83% to 87.81% for 6 test languages). Re-training the system, using the self-labelling approach, yielded a small (i.e., less than 1% absolute) though significant improvement after the first iteration [60].

The image-to-speech results were computed by generating one spoken image caption from each image. This spoken image caption consisted of the segment sequence produced by the image-to-segment transducer, and the resulting speech signal generated by the segment-to-speech transducer. Each test image is matched with the five reference spoken descriptions. The segments generated by the image-to-segment transducer are evaluated using multi-reference BLEU [56]. A similar multi-reference PER is also reported, being the average across all utterances of the minimum, across all five references, of the PER comparing the hypothesis to the reference. The resulting multi-reference PER and BLEU scores are listed in Table IV. Two other scores are also reported: chance and human. Chance is computed by generating a hypothesis exactly the same length as the shortest reference hypothesis, but made up entirely of the most common phone (/n/): the resulting PER is 81.5%. Human BLEU and PER are computed by scoring the human transcriptions against one another: each human transcription was converted to a phone string, and its multi-reference PER and BLEU were computed with respect to the other four human transcriptions. Word-level BLEU scores were not computed, because 1) an unwritten language does not have the concept of a written word; 2) the image-to-speech network has no concept of “words” in the output language.

As Table IV shows, the BLEU scores for the cross-language system are worse than those of the same-language system. The PER scores for the cross-language and same-language image-to-segment transducers are similar though also quite poor. However, as the PER and BLEU scores of the human transcribers show, the task is difficult.

Any two languages will differ in their phoneme set (see [44]. Future research will have to show whether using a different combination of languages yields better results. Initial results on Dutch-to-Mboshi [61] show comparable classification results as Dutch-to-English.

VI. DISCUSSION

This paper investigated whether it is possible to learn speech-to-meaning representations without using text as an intermediate representation, and to test the sufficiency of the learned representations to regenerate speech or translated text, or to retrieve images that depict the meaning of an utterance in an unwritten language. The here-presented results suggest that spoken language human-computer interaction may be possible in an unwritten language. Three types of systems are described: speech-to-translation generation, speech-to-image retrieval, and image-to-speech generation. All three systems use similar neural sequence-to-sequence architectures, and, in fact, re-use many of the same software components.

Mboshi-to-French Example #1
Hyp: j ’ # a i # u n # a b c è s
Ref: j ’ # a i # u n # a b c è s # à # l a # c u i s s e

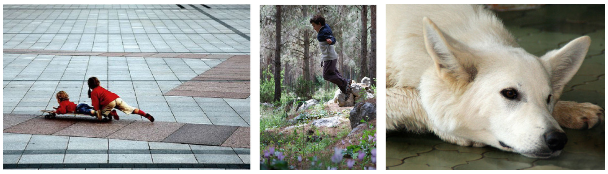
Mboshi-to-French Example #2
Hyp: i l # a # d e # l a # g a l e # p a r t o u t
Ref: i l # m ’ # a # d o n n é # d e # l ’ # e a u # g l a c é e

Fig. 5. Speech-to-translation examples from Mboshi to French. Hyp indicates the hypothesised character sequence in French; Ref indicates the ground truth character sequence French translation; # indicates word boundary.

The speech-to-image retrieval results in Table III are better than the previously published state of the art. Accuracy of our speech-to-translation system (Table II) is worse than the state of the art. Previous papers have reported word-level BLEU scores of up to 7.1 [4] for this task, but it is not clear that small changes in a very small BLEU score adequately characterize differences in the utility of the system for an unwritten language. At this very early stage in technology development for unwritten languages, it may be that analysis of individual examples is the most useful way to characterize areas for future research.

Consider, for example, Fig. 5, which shows two examples generated by the speech-to-translation system from Mboshi to French. Both the hypothesised and ground truth French character sequences are shown. The first example is relatively good: it only misses part of the end of the sentence. The second example shows that the model has difficulty translating a full sentence and diverges to an unconditional language model (unrelated to the source).

Similarly, consider Fig. 6, which shows three examples generated by our image-to-speech system from the validation subset of the FlickR-real corpus, and one image from the evaluation subset. For each image, four transcriptions are shown: two of the five available reference transcriptions (Ref; to give the reader a feeling for the differences among reference transcriptions), the transcription generated by the same-language image-to-speech system (Network), and the transliterations into words (done by hand). The phoneme transcriptions consist of ARPABET phones of [40]. The PER of 70.4% for the same-language system (and 71.7% for the cross-language system) seems to be a pretty high number, until one looks at the examples. The examples show that the system has captured part of the meaning of each image, and that the high PER arises primarily because the neural network chooses to express the meaning of the image using words that differ from those chosen by the human annotators. In particular, note that, although the neural network has no explicit internal representation of words (it simply transduces sub-image sequences into phone sequences), yet, by copying the statistics of its training data onto the generated sentences of the test data, it



FlickR-real 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.

Transliteration: Two men are riding a red and white race.

FlickR-real 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.

Transliteration: A man is jumping in the forest.

FlickR-real Example #3

Ref #1: A close-up of a white dog that is laying its head on its paws.

Ref #2: A large white dog laying on the floor.

Network: SIL +BREATH+ SIL EY B L AE K AE N D AO G Z R AH N IX NG AX B L AE K SIL.

Transliteration: A black an dogs running a black.

Fig. 6. Image examples from the FlickR-real corpus, with for each image, two of its reference transcriptions, the output of the network and its transliteration by the same-language system.

is able to generate outputs that take the form of intelligible and almost-correct image descriptions. In the first two examples, the phone strings shown can be read as English sentences that mislabel boys as men (note that the two captions provided by humans disagree on the gender of the people in the image), but are otherwise almost plausible descriptions of the images. The third example shows several ways in which the network can fail: it has generated a sentence that is syntactically incorrect, and whose semantic content is only partly correct. The phone sequence in this image can be interpreted to contain valid English words, but the transliteration shown here is debatable; since the neural network has no internal representation of "words," it is not clear that transliteration into English words is appropriate in this case. Although the PER and BLEU scores for the cross-language system are lower than those for the same-language system, the results are encouraging.

Due to the lack of text in unwritten languages, standard acoustic models cannot be trained for unwritten languages. In order to train the necessary acoustic models for speech technology in a low-resource language, including unwritten languages, different approaches have been proposed, which can be roughly divided into three strands, each deriving from a different historical tradition within the speech community. First, there is a strand of research deriving from self-organizing speech recognizers. When speech data come without any associated text transcripts, self-organizing systems must create phone-like units directly from the raw acoustic signal while assuming no other information about the language is available, and using these phone-like units to build ASR systems (i.e., the zero resource approach; e.g., [36], [52], [57], [69], [76]). Second, there is a strand of research using the international phonetic alphabet (IPA) to define language-independent phone units for speech technology [66]. Importantly, however, different languages have slightly different productions of each IPA phone (e.g., [34]). Therefore it is necessary to create language-dependent adaptations of each language-independent base phone, which is done through building ASR systems using speech data from multiple languages [43], [66], [70], [71], [74]. The third strand takes its inspiration from the way hearing children learn language and is exemplified by the speech-to-image systems described in the Background section: In addition to the auditory input, hearing children, when learning a language, also have visual information available which guides the language learning process. This third strand compensates the lack of transcribed data with using visual information, from images, to discover word-like units from the speech signal using speech-image associations [2], [23], [27]. Here, we propose to extend or widen this third strand to move beyond going from speech-to-images, to go from speech-to-meaning and from meaning-to-speech. We thus add a new semantic dimension on top of speech and images and that is translated text. We refer to this approach as "unsupervised multi-modal language acquisition".

The goal of the research described in this article was to develop this idea using multi-modal datasets that not only include images but also include translations in a high-resource language (Fig. 1). Parallel data between speech from an unwritten language and translations of that speech signal in another language exist, and additional corpora can fairly easily be collected [7], by field linguists and speech technologists.

Here, the speech-to-meaning and meaning-to-speech approach has been used to discover word-like units from the speech signal using speech-image associations [2], [23], [27]. However, it is possible to push this approach further and searching over subsets of the audio and image can identify sections of audio ("words") that maximally correlate with sections of the image ("objects") [26]. Moreover, unsupervised decomposition of the audio words can be used to deduce phoneme-like units [25].

VII. Conclusion

Three speech technology systems were implemented. The results are encouraging, and suggest that building systems that go directly from speech-to-meaning and from meaning-to-speech, bypassing the need for text, is possible.

This research paves the way for developing speech technology applications for unwritten languages, although more research is needed to build viable systems that can be deployed. The proof-of-concept end-to-end systems we developed were an image-to-speech system, a speech-to-translation system, and a speech-to-image retrieval system. One of our systems outperformed previously reported baselines: an image retrieval system that used multilingual bottleneck features beat the best result reported in the literature for this task.

Speech and language technology systems can be developed for an unwritten language, in a way that is similar to how children learn a language. The speech-to-meaning and meaning-to-speech systems built show that intermediate representations are not necessary to build speech and language technology.

Important avenues for future research are improving the quality of the discovered speech, image and translation encodings, finding the optimal acoustic feature set for the end-to-end systems, and the development of new evaluation metrics that more accurately quantify the utility of a speech technology system in an unwritten language.

ACKNOWLEDGMENT

The authors would like to thank S. Khudanpur and the rest of the Johns Hopkins University team and the local team at Carnegie Mellon University for organizing the JSALT workshop. The work reported here was started at JSALT 2017 in CMU, Pittsburgh.

REFERENCES

[1] G. Adda *et al.*, “Breaking the unwritten language barrier: The BULB project,” in *Proc. Int. Workshop Spoken Lang. Technol. Under-Resourced Lang.*, Yogyakarta, Indonesia, 2016.

[2] A. Alishani, M. Barking, and G. Chrupala, “Encoding of phonology in a recurrent neural model of grounded speech,” in *Proc. Comput. Natural Lang. Learn.*, 2017, pp. 368–378.

[3] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *Proc. Int. Conf. Learn. Representations*, 2015.

[4] S. Bansal, H. Kamper, K. Livescu, A. Lopez, and S. Goldwater, “Pre-training on high-resource speech recognition improves low-resource speech-to-text translation,” 2018. [Online]. Available: https://arxiv.org/abs/1809.01431

[5] A. Bérard, O. Pietquin, C. Servan, and L. Besacier, “Listen and translate: A proof of concept for end-to-end speech-to-text translation,” in *Proc. Neural Inf. Process. Syst. Workshop End-to-End Learn. Speech Audio Process.*, 2016.

[6] L. Besacier, B. Zhou, and Y. Gao, “Towards speech translation of non written languages,” in *Proc. Spoken Lang. Technol. Workshop*, 2006, pp. 222–225.

[7] D. Blachon, E. Gauthier, L. Besacier, G.-N. Kouarata, M. A.-Decker, and A. Rialland, “Parallel speech collection for under-resourced language studies using the LIG-Aikuma mobile device app,” in *Proc. Int. Workshop Spoken Lang. Technol. Under-Resourced Lang.*, Yogyakarta, Indonesia, May 2016.

[8] A. W. Black, “CLUSTERGEN: A statistical parametric speech synthesizer using trajectory modeling,” in *Proc. Int. Conf. Spoken Lang. Process.*, 2006, pp. 1762–1765.

[9] A. W. Black and P. K. Muthukumar, “Random forests for statistical speech synthesis,” in *Proc. Interspeech*, 2015, pp. 1211–1215.

[10] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell,” 2015, *arXiv:1508.01211*.

[11] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, “Large-scale online learning of image similarity through ranking,” *J. Mach. Learn. Res.*, vol. 11, pp. 1109–1135, 2010.

[12] Q. Chen, Z.-H. Ling, and X. Zhu, “Enhancing Sentence embedding with generalized pooling,” 2018, *arXiv:1806.09828*.

[13] J. C. Leavitt, L. Lamel, A. Rialland, M. A.-Decker, and G. Adda, “Corpus based linguistic exploration via forced alignments with a light-weight asr tool,” in *Proc. Lang. Technol. Conf. Human Lang. Technol. Challenge Comput. Sci. Linguistics*, 2017.

[14] J. Deng, K. Li, M. Do, H. Su, and L. F.-Fei, “Construction and analysis of a large scale image ontology,” in *Vision Sci. Soc.*, 2009.

[15] M. A. Di Gangi, R. Dessì, R. Cattoni, M. Negri, and M. Turchi, “Finne-tuning on clean data for end-to-end speech translation: FBK IWSLT 2018,” in *Proc. Int. Workshop Spoken Lang. Transl.*, 2018, pp. 147–152.

[16] L. Duong, A. Anastasopoulos, D. Chiang, S. Bird14, and T. Cohn, “An attentional model for speech translation without transcription,” in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguist.*, 2016, pp. 949–959.

[17] G. Neubig *et al.*, “XNMT: The extensible neural machine translation toolkit,” 2018, *arXiv:1803.00188*.

[18] R. Fer, P. Matejka, F. Grezl, O. Plchot, K. Vesely, and J. H. Černocký, “Multilingually trained bottleneck features in spoken language recognition,” *Comput. Speech Lang.*, vol. 46, pp. 252–267, 2017.

[19] E. Fosler-Lussier, “Contextual word and syllable pronunciation models,” in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 1999.

[20] D. Frossard, “Vgg16 in tensorflow,” 2016, [Online]. Available: https://www.cs.toronto.edu/frossard/post/vgg16/. Accessed: Sep. 14, 2017.

[21] P. Godard *et al.*, “A very low resource language speech corpus for computational language documentation experiments,” 2017, *arXiv:1710.03501*.

[22] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.

[23] D. Harwath and J. Glass, “Deep multimodal semantic embeddings for speech and images,” in *Proc. Workshop Autom. Speech Recognit. Understanding*, Scottsdale, Arizona, USA, 2015, pp. 237–244.

[24] D. Harwath, G. Chuang, and J. Glass, “Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech,” in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 4969–4973.

[25] D. Harwath and J. Glass, “Towards visually grounded sub-word speech unit discovery,” 2019. [Online]. Available: https://arxiv.org/pdf/1902.08213.pdf

[26] D. Harwath, A. Recasens, D. Suris, G. Chuang, A. Torralba, and J. Glass, “Jointly discovering visual objects and spoken words from raw sensory input,” in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 649–665.

[27] D. Harwath, A. Torralba, and J. Glass, “Unsupervised learning of spoken language with visual context,” in *Proc. Adv. Neural Inform. Process. Syst.*, 2016, pp. 1858–1866.

[28] M. Hasegawa-Johnson, A. Black, L. Ondel, O. Scharenborg, and F. Ciannella, “Image2speech: Automatically generating audio descriptions of images,” in *Proc. Int. Conf. Natural Lang., Signal Speech Process., Casablanca, Morocco*, 2017.

[29] M. Hasegawa-Johnson, A. Black, L. Ondel, O. Scharenborg, and F. Ciannella, “Image2speech: Automatically generating audio descriptions of images,” *J. Int. Sci. Gen. Appl.*, vol. 1, pp. 19–27, 2018.

[30] M. Hasegawa-Johnson *et al.*, “ASR for under-resourced languages from probabilistic transcription,” *IEEE/ACM Trans. Audio, Speech Lang.*, vol. 25, no. 1 pp. 46–59, Jan. 2017.

[31] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.

[32] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, “The ATIS spoken language systems pilot corpus,” in *Proc. HLT ’90 Proc. Workshop Speech Natural Lang.*, 1990, pp. 96–101.

[33] M. Hodosh, P. Young, and J. Hockenmaier, “Framing image description as a ranking task: Data, models and evaluation metrics,” *J. Artif. Intell. Res.*, vol. 47, pp. 853–899, 2015.

[34] P.-S. Huang and M. Hasegawa-Johnson, “Cross-dialectal data transferring for gaussian mixture model training in arabic speech recognition,” in *Proc. Int. Conf. Arabic Lang. Process. (CITALA), Rabat, Morocco*, 2012, pp. 119–122.

[35] H. Inaguma, X. Zhang, Zh. Wang, A. Renduchintala, S. Watanabe, and K. Duh, “The JHU/KyotoU speech translation system for IWSLT 2018,” in *Proc. Int. Workshop Spoken Lang. Transl.*, 2018, pp. 153–159.

[36] A. Jansen *et al.*, “A summary of the 2012 JH CLSP Workshop on zero resource speech technologies and models of early language acquisition,” in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 8111–8115.

[37] F. Jelinek, “Continuous speech recognition by statistical methods,” *Proc. IEEE*, vol. 64, no. 4, pp. 532–556, Apr. 1976.

[38] F. Jelinek, L. R. Bahl, and R. L. Mercer, "The design of a linguistic statistical decoder for the recognition of continuous speech," *IEEE Trans. Inform. Theory*, vol. 21, no. 3, pp. 250–256, May 1975.

[39] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Seattle, Washington, USA, Oct. 2013, pp. 1700–1709.

[40] 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 *Proc. Int. Workshop Spoken Lang. Transl.*, 2014, pp. 73–79.

[41] S. Krauwer, "The basic language resource kit (BLARK) as the first milestone for the language resources roadmap," in *Proc. Int. Workshop Speech Comput., Moscow, Russia*, 2003, pp. 8–15.

[42] K. Lenzo, "The CMU pronouncing dictionary," 2014. [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[43] J. Lööf, C. Gollan, and H. Ney, "Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a polish speech recognition system," in *Proc. Interspeech*, 2009.

[44] I. Maddieson, *Patterns of Sounds*. Cambridge, U.K.: Cambridge Univ. Press, 1984.

[45] F. Malfrere and T. Dutoit, "High-quality speech synthesis for phonetic speech segmentation," in *Proc. Eurospeech*, 1997, pp. 2631–2634.

[46] P. K. Muthukumar and A. W. Black, "Automatic discovery of a phonetic inventory for unwritten languages for statistical speech synthesis," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 2594–2598.

[47] B. Nash-Webber, "Semantic support for a speech understanding system," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 23 no. 1: pp. 124–129, Feb. 1975.

[48] G. Neubig, "XNMT," [Online]. Available: https://github.com/neulab/xnmt/

[49] G. Neubig, "Neural machine translation and sequence-to-sequence models: A tutorial," 2017, *arXiv:1703.01619*.

[50] J. Niehues, R. Cattoni, S. Stüker, M. Cettolo, M. Turchi, and M. Federico, "The IWSLT 2018 evaluation campaign," in *Proc. Int. Workshop Spoken Lang. Transl.*, 2018, pp. 2–6.

[51] C. K. Ogden and I. A. Richards, *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*. Orlando, FL, USA: Harcourt Brace Jovanovitch, 1923.

[52] L. Ondel, L. Burget, and J. Černocký, "Variational inference for acoustic unit discovery," in *Proc. Comput. Sci.*, 2016, pp. 80–86.

[53] N. Oostdijk *et al.*, "Experiences from the spoken Dutch Corpus Project," in *Proc. Int. Conf. Lang. Resour. Eval., Las Palmas de Gran Canaria*, 2002, pp. 340–347.

[54] M. Ostendorf and S. Roukos, "A stochastic segment model for phoneme-based continuous speech recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 12, pp. 1857–1869, Dec. 1989.

[55] J.-Y. Pan, H.-J. Yang, P. Duygulu, and C. Faloutsos, "Automatic image captioning," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2004, pp. 1987–1990.

[56] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguist.*, 2002, pp. 311–318.

[57] A. S. Park and J. R. Glass, "Unsupervised pattern discovery in Speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16 no. 1: pp. 186–197, Jan. 2008.

[58] L. Plaza, E. Lloret, and A. Aker, "Improving automatic image captioning using text summarization techniques," in *Proc. Int. Conf. Text, Speech Dialogue*, 2010, pp. 165–172.

[59] O. Scharenborg *et al.*, "Linguistic unit discovery from multi-modal inputs in unwritten languages: Summary of the 'Speaking Rosetta' JSALT 2017 Workshop," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Calgary, Alberta, Canada, Apr. 2018, pp. 4979–4983.

[60] O. Scharenborg *et al.*, "Building an ASR system for a low-resource language through the adaptation of a high-resource language ASR system: Preliminary results," in *Proc. ICNLSSP*, Casablanca, Morocco, 2017.

[61] O. Scharenborg, P. Ebel, F. Ciannella, M. Hasegawa-Johnson, and N. Dehak, "Building an ASR system for mboshi using a cross-language definition of acoustic units approach," in *Proc.Workshop Spoken Lang. Technologies Under-Resourced Lang.*, 2018.

[62] J. Shen *et al.*, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 4779–4783.

[63] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image classification," 2014. [Online]. Available: https://arxiv.org/abs/1409.1556, Accessed: Sep. 14, 2017.

[64] S. Stüker *et al.*, "Innovative technologies for under-resourced language documentation: The Bulb project," in *Proc. Collaboration Comput. Under-Resourced Lang.*, Portorozz~ Slovenia, 2016.

[65] I. Sutskever, O. Vinyals, and Q. V Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[66] A. Waibel, T. Schultz, "Experiments on cross-language acoustic modelling," in *Proc. Interspeech*, 2001.

[67] R. Teranishi and N. Umeda, "Use of pronouncing dictionary in speech synthesis experiments," in *Proc. Int. Congr. Acoust.*, Tokyo, Japan, Aug. 1968, pp. B–5–2.

[68] N. Umeda, "Linguistic rules for text-to-speech synthesis," *Proc. IEEE*, vol. 64 no. 4, pp. 443–451, Apr. 1976.

[69] B. Varadarajan, S. Khudanpur, and E. Dupoux, "Unsupervised learning of acoustic sub-word units," in *Proc. ACL Human Lang. Technol.: Short Papers*, 2008, pp. 165–168.

[70] K. Vesely, M. Karafiát, F. Grezl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Proc. Spoken Lang. Technol. Workshop*, 2012, pp. 336–341.

[71] N. T. Vu, F. Metze, and T. Schultz, "Multilingual bottle-neck features and its application for under-resourced languages," in *Proc. 3rd Workshop Spoken Lang. Technol. Under-Resourced Lang.*, Cape Town, S. Africa, May 2012. MICA.

[72] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly transcribe foreign speech," 2017, *arXiv:1703.08581*.

[73] Y. Wu, M. Schuster, Z. Chen, and Q. Le, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*.

[74] H. Xu, V. H. Do, X. Xiao, and E. S. Chng, "A comparative study of BNF and DNN multilingual training on cross-lingual low-resource speech recognition," in *Proc. Interspeech*, 2015, pp. 2132–2136.

[75] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 4651–4659.

[76] Y. Zhang and J. R. Glass, "Towards multi-speaker unsupervised speech pattern discovery," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2010, pp. 4366–4369.