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, Emmanuel Dupoux, *Senior Member, IEEE*

***Abstract*—Speech technology plays an important role in our everyday life. Speech is, among others, used for human-computer interaction, including, for instance, information retrieval and online 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 paper 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.**

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], [35]. 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.

OS is with the Multimedia Computing Group, Delft University of Technology, the Netherlands (part of this work was carried out while she was with the Centre for Language Studies, Radboud University Nijmegen, the Netherlands). LB is with LIG - Univ Grenoble Alpes (UGA), France. AB, FM, GN, SP, and PA are with Carnegie Mellon University, Pittsburgh, PA, U.S.A., FC was with Carnegie Mellon University, Pittsburgh, USA. MHJ and LW are with the Beckman Institute, University of Illinois, Urbana-Champaign, USA. SS and MM are with Karlsruhe Institute of Technology, Germany. PG is with LIMSI, Paris, France. LO is with Brno University, Czech Republic and Johns Hopkins University, Baltimore, MD, U.S.A. MD, EL, RR, and ED are with ENS/CNRS/EHESS/INRIA, Paris, France.

Manuscript received August 8, 2019.

Speech technology is typically viewed as an X-to-speech or speech-to-X task, where X is text. A crucial component of any speech technology system is the acoustic phone(me) model set. In speech-to-text (i.e., automatic speech recognition) systems, the acoustic models are trained using (speech,phoneme transcription) pairs, while during testing, the acoustic phoneme models are used to find the optimal sequence of words by aligning sequences of acoustic models, determined on the basis of the phoneme transcription of the words, with the speech signal. In text-to-speech (speech synthesis) systems, the acoustic models are used to generate the pronunciation of a word by sequencing the acoustic models of the phone(me) transcription of the word. In the case of an unwritten language, the X cannot be text, and thus needs to be redefined. Here, we propose to learn mappings from speech to meaning, and from meaning to speech, directly, without using text as an intermediate representation, in order to build speech technology for unwritten languages.

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 [27] 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. Consider the model of semantics shown in Fig. 1. In this model, the logical propositional form of an utterance's meaning is unknown, but instead, we have two or three different instantiations of the same meaning: a speech signal in one language matched to a text translation in another language, or matched to an image that depicts the situation that the proposition describes. 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 the speech, the translation, and/or retrieve the image from a database?

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], [58]. 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

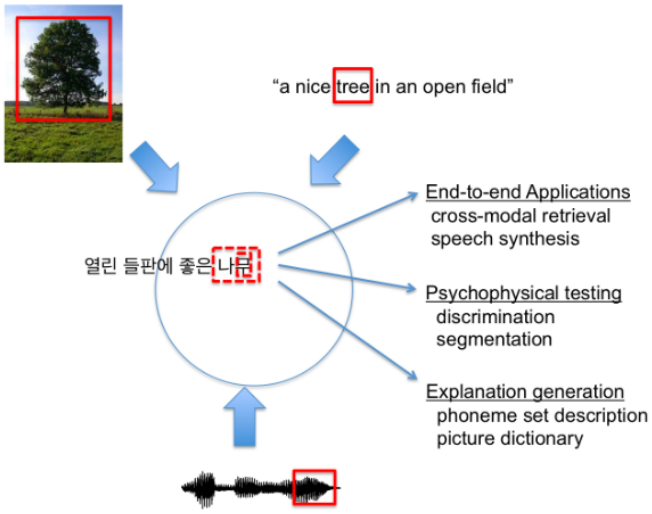

Fig. 1: A model of semantics for speech technology development in an unwritten language. The speech signal (bottom of the figure) has some propositional content which is unknown and not directly observable (represented by the Korean sentence in the center of the figure). Instead of directly observing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

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], [18], [22], 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, arguably similar to how children acquire their first language. 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 [23], [24], 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 building of traditional acoustic models. The image-to-speech application creates acoustic models on the basis of automatically discovered speech units.

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 V 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 VI gives experimental results, Section VII is discussion, and Section VIII 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 [14], [5]. 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 [40]. 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 [13], [30]. The E2E systems were however outperformed by the pipelined system: [30] 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 [13].

The image-to-speech generation task was proposed in [23], [24], 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 [43]. The task of generating keywords from an image led to alternate definitions using text summarization techniques [46] and image-to-text retrieval techniques [28]. 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 [60].

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,

[1]Note, a summary and initial results of this work were presented in [47], also available in the HAL repository: 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 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 [23], [24].

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 [18]. 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) [22].

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 [19]. Searching over subsets of the audio and image can identify sections of audio ("words") that maximally correlate with sections of the image ("objects") [21]; unsupervised decomposition of the audio words can be used to deduce phoneme-like units [20]. It is possible to use a relatively uniform convolutional architecture across all three modalities [22], but performance improvements are possible by using a gated recurrent highway unit for the speech encoder [2].

III. Architecture

Figure 2 shows the schematic of a DNN-based system that maps from speech, translated text, or images into a hidden semantic space, and then regenerates speech or text from the underlying representation. All three input modalities are projected into a common semantic encoding, using an encoding network that uses a combination of convolutional and recurrent layers, as described later in this section. The input to the encoding network is different, depending on the modality. 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, with weights pre-trained for the ImageNet image classification task by [16]. 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.

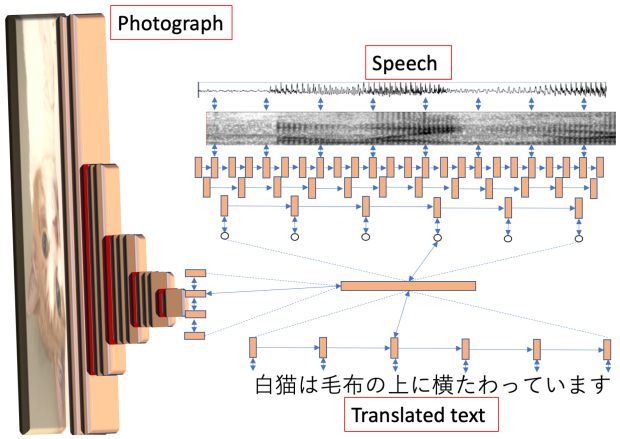

Fig. 2: Proposed neural architecture. Separate encoder and decoder networks are trained for each of the three modalities. The figure shows the speech encoder (a pyramidal LSTM), and decoders (LSTMs with attention-weighted input context vectors) that would generate an image output or a translated text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

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 similarity 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 Figure 2 represents the translation f_{YX}, from any modality X to any other modality Y, as the composition of an encoder g_X and a decoder h_Y, thus for example,

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

The encoder, g_X, is modeled as a pyramidal long-short term memory network (pyramidal LSTM): a three-layer LSTM 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). These encoders have been successfully used in speech recognition [10].

Let $\vec{e}_{l,t}$ be the t^{th} encoder state vector at level l of the network. Each state vector is computed by applying a memory gate to the preceding state vector of the same layer, and an input gate to the concatenation of two consecutive state vectors from the layer below; let us represent these operations by the nonlinear function γ, thus

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

The input to the encoder is the matrix of modality-dependent feature vectors, e.g., for speech input, $\vec{e}_{0,t} = \vec{x}_t$. The output is a sequence of encoder state vectors at the L^{th} level,

$$g_X(X) = [\vec{e}_{L,1}, \ldots, \vec{e}_{L,D_X}], \quad (3)$$

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

The decoder, h_Y, is modeled as a sequence-to-sequence neural decoder with attention-weighted inputs. Let $\hat{Y} = [\hat{y}_1, \ldots, \hat{y}_{T_Y}]$ be the sequence generated by the decoder h_Y. Each output character or phone, $\hat{y}_i$, is generated from a decoder state vector, $\vec{s}_i$. The decoder state vector sequence, in turn, is generated from a set of attention-weighted context vectors, $\vec{c}_i$, computed from the encoder state vectors as

$$\vec{c}_i = \sum_{t=1}^{D_X} a_{it} \vec{e}_{L,t}, \tag{4}$$

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

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

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

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

The probability of a sequence of output symbols is computed by a softmax transformation, with weight vectors $\vec{w}_k$, of an input composed of the concatenated context vector and decoder state vector, thus the probability of generating the j^{th} symbol type in the i^{th} output slot is

$$P(\hat{y}_i i = j) = \frac{\exp(\vec{w}_j^T [\vec{s}_i, \vec{c}_i])}{\sum_k \exp(\vec{w}_k^T [\vec{s}_i, \vec{c}_i])}. \tag{7}$$

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.

Two types of loss functions were used in this work. Speech-to-translation was trained in order to minimize cross-entropy between the reference and hypothesis character sequences. Cross-entropy was also used to train the image-to-speech generation system. To that end, we first created an equivalence between phone symbols and sequences of cepstral vectors, using a combination of Kaldi and Festvox as described below in Section V-C. In both cases, the cross-entropy loss can be written as

$$\mathcal{L}\left(Y, f_{YX}(X)\right) = -\sum_{i=1}^{T_Y} \ln P(\hat{y}_i = y_i) \tag{8}$$

The speech-to-image retrieval task requires us to measure the similarity between two vector sequences, $g_X(X)$ and $g_Z(Z)$, of different lengths D_X and D_Z. Following [18], this is performed by choosing one of the modalities as the reference (X, say), and the other as the target (Z, say), and then finding, for each reference vector, the best-matching target vector:

$$\mathcal{L}\left(g_X(X), g_Z(Z)\right) = -\sum_{t=1}^{D_X} \max_{1 \le i \le D_Z} \frac{\vec{e}_{L,t}(X) \cdot \vec{e}_{L,i}(Z)}{\|\vec{e}_{L,t}(X)\| \cdot \|\vec{e}_{L,i}(Z)\|} \tag{9}$$

IV. Data

In order to train a neural network for speech-to-translation generation, speech-to-image retrieval, or image-to-speech generation, it is necessary to have a training corpus with matched pairs of spoken utterances, text translations, and/or images. For our experiments, we used data in one language that is truly unwritten (i.e., it has no standard system of orthography), in one simulated unwritten language (a language that has a written form, but whose written form was not included in the training corpus), and in three written languages. The unwritten language we used is Mboshi, which is a Bantu language (Bantu C25) of Congo-Brazzaville [1], [51]. Mboshi was chosen as a test language because Mboshi utterances and their paired French translations were available to us through the BULB project [1]. We did not have available to us a three-way corpus of matched images, utterances, and translations of sentences of a truly unwritten language, so instead, we used the FlickR-real corpus of English utterances matched to images and Japanese text translations; thus English served as a simulated unwritten language. The three written languages used in these experiments were French and Japanese (the translation targets for the Mboshi and FlickR-real corpora, respectively) and Dutch (a source language used to define phones for the English language image-to-speech task).

Table I gives an overview of the characteristics of the multi- and unimodal datasets, which were used in the experiments. The Mboshi corpus [17] was collected using a real language documentation scenario, using ligaikuma,[2] a recording application for language documentation [7]. The Mboshi corpus is a multilingual corpus consisting of 5k speech utterances (approximately 4 hours of speech) in Mboshi with hand-checked French text translations. Additionally, the corpus contains linguists' transcriptions in a non-standard graphemic form which is rather close to the phonology of the language [1], [17]. The corpus is augmented with automatic forced-alignments between the Mboshi speech and the linguists' transcriptions of the Mboshi speech in phonemes [31]. The corpus and forced alignments are made available to the research community.[3]

The FlickR-real speech database is a tri-modal (speech, translated text, images) corpus. The images in this dataset were selected through user queries for specific objects and actions from the FlickR photo sharing website [28]. Each image contains five descriptions in natural language which were collected using a crowdsourcing platform (Amazon Mechanical Turk; AMT). AMT was also used by [18] to obtain 40K spoken versions of the captions. These are made available online.[4] We augmented this corpus by adding Japanese translations (Google MT) for all 40K captions, as well as Japanese tokenization.

The Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN, [42]) is a corpus of almost 9M words of Dutch spoken in the Netherlands and in Flanders (Belgium) in over 14

[2]http://lig-aikuma.imag.fr

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

[4]https://groups.csail.mit.edu/sls/downloads/flickraudio/

TABLE I: Overview of the databases.

Data set	Language	Size	Aligned translations	Aligned images	#spkrs
Mboshi	Mboshi	5h	yes (French - Human)	no	3
FlickR-real speech	English	62h	yes (Japanese - MT)	yes	183
Corpus Spoken Dutch	Dutch	64h	no	no	133

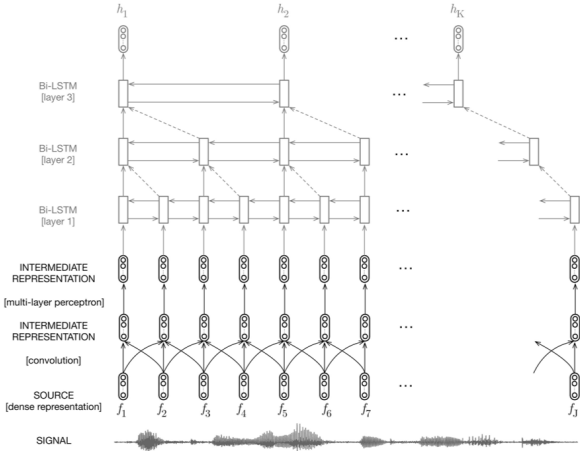

Fig. 3: The encoder architecture for speech-to-translation experiments was a three-layer bi-directional pyramidal LSTM, observing speech features computed by a one-layer convolutional network over the top of MFCCs.

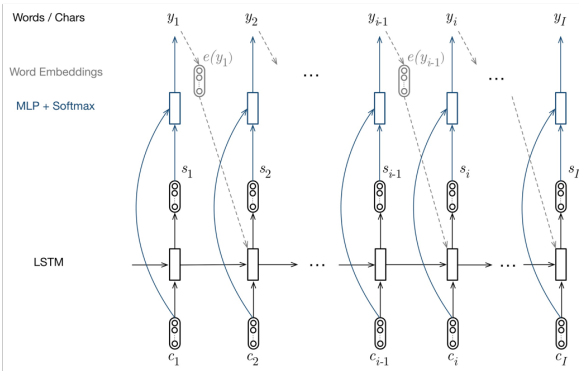

Fig. 4: The decoder architecture for speech-to-translation experiments was a one-layer LSTM generating characters as output (word outputs were also tested, but were not as successful).

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.

V. Experimental set-up

A. Speech-to-translation

We built end-to-end speech-to-translation systems with the neural sequence-to-sequence machine translation toolkit XNMT [38], [15] 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 [33], [52], [3], [39] of XNMT.

The speech encoder for 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 Mel-frequency cepstral coefficients (MFCCs) inputs.

The decoder, shown in Figure 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 one-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 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.

B. Speech-to-Image

The speech-to-image system was build using PyTorch, and trained and tested on 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 PyTorch system consisted of an image encoder using visual features extracted using a pretrained ResNet-152 [26] with the top layer removed. These features were then fed into a fully connected layer with 1024 units. The speech was encoded using a 1d 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 multi-dimensional attention layer [11]. The resulting embeddings were normalized to have unit L2 norm, and used the cosine similarity score (Eq. 9) between the image and speech embeddings to perform the retrieval task.

Two types of acoustic features were compared: 1) MFCCs (baseline features), similar to [18] 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). 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. 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 [24] consists of four types of standard open-source software toolkits: 1) a VGG16 visual object recognizer which converts each image into a sequence of feature vectors. 2) XNMT, which accepts image feature vectors as inputs, and generates speech unit sequences as output. 3) A text-to-speech (TTS) system, Clustergen [8], which generates audio from each speech unit sequence. 4) A system that discovers discrete speech units in the unwritten languages. These discovered speech units are used to create a transcription of the speech, which is needed to train the TTS system. The image-to-speech system was trained on the FlickR-real corpus using (image,sequence) pairs, where the sequence consists of a sequence of speech units. For training, 6000 training images were used. When an image is part of the training or validation corpus, all of its captions are used, thus the FlickR-real training corpus included 30,000 image-audio pairs (6000 distinct images). The validation set consisted of 1000 validation images, while a further 1000 images were used for testing.

Two image-to-speech systems were created. In the first system, the sequence of speech units was obtained using a forced alignment of the FlickR-real data obtained with Kaldi. This system was built in order to determine the upper-bound performance of such an image-to-speech system. The second system is the system as it would and could be employed in the setting of an unwritten language. Here, the speech units were automatically discovered from the speech signal (see below).

1) The VGG16 visual object recognizer: The VGG16 object recognizer was the TensorFlow re-implementation, by [16], of the best single network solution [50] 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 [12].

2) XNMT architecture: The image-to-speech 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, pyramidal LSTM, with a 128-dimensional state vector. The attender is a three-layer perceptron. 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 a context vector 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.

3) TTS-system Clustergen: The text-to-speech (TTS) system used is 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, and a machine learning algorithm (e.g., regression tree [8] or random forest [9]) is applied to learn the mapping. In other words, 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 [37], which makes it suitable for our low-resource scenario.

The input to Clustergen is a waveform file plus symbolic sequences of speech units; the output is a simple synthesizer and a Mel-cepstral distortion measure [54] on held out data.

4) Cross-language definition of speech units: In the cross-language definition of units approach [48], [49] a DNN was trained on a high-resource language, Dutch, which was subsequently mapped to English (of the FlickR-real database). Since the phoneme inventories of Dutch and English are different, the Dutch phonemes that do not exist in English are removed by removing the corresponding vectors from the soft-max layer. The phones that exist in English but not in Dutch were created through a linear extrapolation between existing Dutch acoustic units in the soft-max layer using:

$$\vec{V}_{|\phi|,L2} = \vec{V}_{|\phi|,L1:1} + \alpha(\vec{V}_{|\phi|,L1:2} - \vec{V}_{|\phi|,L1:3})$$

where $\vec{V}_{|\phi|,L2}$ is the vector of the missing L2 phone $\phi, L2$ that needs to be created, $\vec{V}_{|\phi|,L1:x}$ are the vectors of the Dutch L1 phones $\phi, L1:x$ in the soft-max layer that are used to create the vector for the missing English phone $\phi, L2$. Among the three Dutch phones, L1:1 refers to the phone which is used as the starting point from which to extrapolate the missing L2 phone, and L1:2 and L1:3 refer to the L1 phones whose displacement is used as an approximation of the displacement between the Dutch L1 vector and the L2 phone that should be created. α is a factor corresponding to the approximation of the displacement of $\vec{V}_{|\phi|,L2}$ from $\vec{V}_{|\phi|,L1:1}$.

The acoustic units that are used to initialize the new English acoustic feature vectors are chosen on the basis of their linguistic similarity to the English phonemes which need to be created. Subsequently, the acoustic units are iteratively retrained using their self-labels. This adapted system was then created to generate a phone transcription of the FlickR-real target data, which was used to train the speech synthesis system.

VI. Results

A. Speech-to-translation

The speech-to-translation system was trained and tested for 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

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

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**

scores (average recall accuracy of character 1-gram through 5-gram sequences [44]) 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].

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 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, the MBN features are superior to the MFCC features, and show state of the art results, 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

In order to train the image-to-speech system, speech was first segmented using either an English forced-alignment system, or a cross-language Dutch-to-English speech recognizer. 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 ASR systems (e.g., [25] 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.

The image-to-speech results were computed by generating one spoken image caption from each image, computing its PER and phone-level BLEU score with respect to each of the five reference captions for the same image, and then averaging. The resulting average PER and BLEU scores are listed in Table IV. One baseline and one upper-bound score

TABLE IV: Image-to-speech results (Phone-level BLEU scores and phone error rates (PER (%)) on the val(idation) and test sets of the upper-bound and the simulated unwritten language systems. * indicates a PER <chance (Student's T, Chebyshev standard error, p <0.001; chance=90.2%)

System	Val BLEU	Val PER	Test BLEU	Test PER
Human Transcriptions			27.0	88.0
Upper-bound	13.7	87.9	13.7	84.9*
Unwritten	5.4	115	6.1	101
Chance				90.2

are also listed. The baseline score is chance accuracy, which is computed by generating a hypothesis exactly the same length as the correct hypothesis, but made up entirely of the most common phone (/n/): the resulting PER is 90.2%. The upper-bound score is the error rate of the human transcriptions, scored against one another: each human transcription was converted to a phone string, and the pairwise differences between the five human-generated phone strings were scored in terms of phone-level BLEU and PER. 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 PER and BLEU scores for the unwritten language system are quite poor. However, as the PER and BLEU scores of the upper-bound system and the human transcriptions show, the task is difficult. The average PER of human transcriptions of the spoken image descriptions, scored against one another, is 88%; an image-to-speech system trained using the human transcriptions achieves an average PER with respect to the human transcriptions that is only slightly better, at 84.9%. This PER is however significantly better than chance.

VII. 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.

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 (as we obtained a word-level BLEU score of around zero, which was not reported, we only reported character-level BLEU scores), but considering that the state of the art for the Mboshi corpus is a word-level BLEU score of only 7.1 [4], it is possible that word-level BLEU is an inappropriate measure for evaluating such systems. There is reason to believe that the low BLEU scores dramatically

under-estimate the utility of these systems, and that further research is necessary in order to define evaluation metrics that adequately measure the utility of speech technology in an unwritten language. Consider, for example, Fig. 5, which shows two examples generated by our image-to-speech system from the validation subset of the FlickR-real corpus. 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 upper-bound image-to-speech system (Network), and the transliterations into words (done by hand). The phoneme transcriptions consist of ARPABET phones of [34]. The difference between 84.9% PER for the upper-bound system and chance PER is statistically significant, but 84.9% error still seems to be a pretty high number, until one looks at the examples. The examples show that the system has captured most 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 is able to generate outputs that take the form of intelligible and almost-correct image descriptions. In these 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.

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., [32], [41], [55], [45], [61]). Second, there is a strand of research using the international phonetic alphabet (IPA) to define language-independent phone units for speech technology [53]. Importantly, however, different languages have slightly different productions of each IPA phone (e.g., [29]). 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 [53], [36], [57], [56], [59]. 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 [18], [2], [22]. 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".

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.*

Fig. 5: 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 upper-bound system.

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 (Figure 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.

VIII. Conclusions

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 work reported here was started at JSALT 2017 in CMU, Pittsburgh, and was supported by JHU and CMU via grants from Google, Microsoft, Amazon, Facebook, and Apple. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by NSF grant number OCI-1053575. Specifically, it used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC).

OS was partially supported by a Vidi-grant from NWO (276-89-003) and partially by a Delft Technology Fellowship from Delft University of Technology. PG, MM and SS were funded by the French ANR and the German DFG under grant ANR-14-CE35-0002 (BULB project). MD, EL, RR and ED were funded by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), and ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL*.

The authors would like to thank Sanjeev Khudanpur and the rest of the Johns Hopkins University team and the local team at Carnegie Mellon University for organizing the JSALT workshop.

REFERENCES

[1] Gilles Adda, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitri Idiatov, Guy-Noël Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Annie Rialland, Mark Van de Velde, François Yvon, and Sabine Zerbian. Breaking the unwritten language barrier: The BULB project. In *Proceedings of SLTU*, Yogyakarta, Indonesia, 2016.

[2] Afra Alishani, Marie Barking, and Grzegorz Chrupala. Encoding of phonology in a recurrent neural model of grounded speech. In *Proceedings Computational Natural Language Learning (CoNLL)*, pages 368–378, 2017.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015.

[4] Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. Downloaded 4/19/2019 from https://arxiv.org/abs/1809.01431, 2018.

[5] Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NIPS Workshop on end-to-end learning for speech and audio processing*, 2016.

[6] Laurent Besacier, Bowen Zhou, and Yuqing Gao. Towards speech translation of non written languages. In *Spoken Language Technology Workshop, 2006. IEEE*, pages 222–225, 2006.

[7] David Blachon, Elodie Gauthier, Laurent Besacier, Guy-Noël Kouarata, Martine Adda-Decker, and Annie Rialland. Parallel speech collection for under-resourced language studies using the LIG-Aikuma mobile device app. In *Proceedings of SLTU*, Yogyakarta, Indonesia, May 2016.

[8] Alan W. Black. CLUSTERGEN: A statistical parametric speech synthesizer using trajectory modeling. In *Proceedings of ICSLP*, pages 1762–1765, 2006.

[9] Alan W. Black and Prasanna Kumar Muthukumar. Random forests for statistical speech synthesis. In *Proceedings of Interspeech*, pages 1211–1215, 2015.

[10] William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*, 2015.

[11] Qian Chen, Zhen-Hua Ling, and Xiaodan Zhu. Enhancing Sentence Embedding with Generalized Pooling. 2018.

[12] 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.

[13] Mattia Antonino Di Gangi, Roberto Dessì, Roldano Cattoni, Matteo Negri, and Marco Turchi. Finne-tuning on clean data for end-to-end speech translation: FBK IWSLT 2018. In *International Workshop on Spoken Language Translation*, pages 147–152, 2018.

[14] Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird14, and Trevor Cohn. An attentional model for speech translation without transcription. In *Proceedings of NAACL-HLT*, pages 949–959, 2016.

[15] Graham Neubig et al. Xnmt: The extensible neural machine translation toolkit. In *arXiv:1803.00188, 2018*, 2018.

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

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

[18] D. Harwath and J. Glass. Deep multimodal semantic embeddings for speech and images. In *Proceedings of ASRU*, pages 237–244, Scottsdale, Arizona, USA, 2015.

[19] David Harwath, Galen Chuang, and James Glass. Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech. In *Proceedings of ICASSP*, pages 4969–4973, 2018.

[20] David Harwath and James Glass. Towards visually grounded sub-word speech unit discovery. Downloaded 4/19/2019 from https://arxiv.org/pdf/1902.08213.pdf, 2019.

[21] David Harwath, Adria Recasens, Didac Suris, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *Proceedings European Conference on Computer Vision (ECCV)*, pages 649–665, 2018.

[22] David Harwath, Antonio Torralba, and James Glass. Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems*, pages 1858–1866, 2016.

[23] Mark Hasegawa-Johnson, Alan Black, Lucas Ondel, Odette Scharenborg, and Francesco Ciannella. Image2speech: Automatically generating audio descriptions of images. In *Proceedings of ICNLSSP, Casablanca, Morocco*, 2017.

[24] Mark Hasegawa-Johnson, Alan Black, Lucas Ondel, Odette Scharenborg, and Francesco Ciannella. Image2speech: Automatically generating audio descriptions of images. In *Journal of International Science and General Applications*, volume 1, pages 19–27, 2018.

[25] Mark Hasegawa-Johnson, Preethi Jyothi, Daniel McCloy, Majid Mirbagheri, Giovanni di Liberto, Amit Das, Bradley Ekin, Chunxi Liu, Vimal Manohar, Hao Tang, Edmund C. Lalor, Nancy Chen, Paul Hager, Tyler Kekona, Rose Sloan, , and Adrian KC Lee. Asr for under-resourced languages from probabilistic transcription. *IEEE/ACM Trans. Audio, Speech and Language*, 25(1):46–59, 2017.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*. IEEE, 2016.

[27] Charles T. Hemphill, John J. Godfrey, and George R. Doddington. The ATIS spoken language systems pilot corpus. In *HLT '90 Proceedings of tne Workshop on Speech and Natural Language*, pages 96–101, 1990.

[28] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[29] P.-S. Huang and M. Hasegawa-Johnson. Cross-dialectal data transferring for gaussian mixture model training in arabic speech recognition. In *International Conference on Arabic Language Processing (CITALA), Rabat, Morocco*, pages 119–122, 2012.

[30] Hirofumi Inaguma, Xuan Zhang, Zhiqi Wang, Adithya Renduchintala, Shinji Watanabe, and Kevin Duh. The JHU/KyotoU speech translation system for IWSLT 2018. In *International Workshop on Spoken Language Translation*, pages 153–159, 2018.

[31] Annie Rialland Martine Adda-Decker Gilles Adda Jamison Cooper-Leavitt, Lori Lamel. Corpus based linguistic exploration via forced alignments with a light-weight asr tool. In *Language and Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics*, 2017.

[32] Aren Jansen et al. A summary of the 2012 JH CLSP Workshop on zero resource speech technologies and models of early language acquisition. In *Proceedings of ICASSP*, 2013.

[33] Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[34] Kevin Kilgour, Michael Heck, Markus Müller, Matthias Sperber, Sebastian Stüker, and Alex Waibel. The 2014 kit iwslt speech-to-text systems for english, german and italian. In *International Workshop on Spoken Language Translation (IWSLT)*, pages 73–79, 2014.

[35] S. Krauwer. The basic language resource kit (blark) as the first milestone for the language resources roadmap. In *Proc. International Workshop Speech and Computer, Moscow, Russia*, pages 8–15, 2003.

[36] J. L??f, C. Gollan, and H. Ney. Cross-language bootstrapping for unsupervised acoustic model training: rapid development of a polish speech recognition system. In *Proceedings of Interspeech*, 2009.

[37] Prasanna Kumar Muthukumar and Alan W. Black. Automatic discovery of a phonetic inventory for unwritten languages for statistical speech synthesis. In *Proceedings of ICASSP*, 2014.

[38] Graham Neubig. Xnmt. https://github.com/neulab/xnmt/.

[39] Graham Neubig. Neural machine translation and sequence-to-sequence models: A tutorial. *arXiv preprint arXiv:1703.01619*, 2017.

[40] J. Niehues, R. Cattoni, S. Stüker, M. Cettolo, M. Turchi, and M. Federico. The IWSLT 2018 evaluation campaign. In *International Workshop on Spoken Language Translation*, pages 2–6, 2018.

[41] Lucas Ondel, Lukáš Burget, and Jan Černocký. Variational inference for acoustic unit discovery. In *Procedia Computer Science*, pages 80–86, 2016.

[42] Nelleke Oostdijk, W. Goedertier, F. Van Eynde, Lou Boves, Jean-Pierre Martens, Michael Moortgat, and Harald Baayen. Experiences from the spoken dutch corpus project. In *Proceedings of LREC, Las Palmas de Gran Canaria*, pages 340–347, 2002.

[43] Jia-Yu Pan, Hyung-Jeong Yang, Pinar Duygulu, and Christos Faloutsos. Automatic image captioning. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2004.

[44] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[45] Alex S. Park and James R. Glass. Unsupervised Pattern Discovery in Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):186–197, 2008.

[46] Laura Plaza, Elena Lloret, and Ahmet Aker. Improving automatic image captioning using text summarization techniques. In *International Conference on Text, Speech and Dialogue*, pages 165–172, 2010.

[47] Odette Scharenborg, Laurent Besacier, Alan Black, Mark Hasegawa-Johnson, Florian Metze, Graham Neubig, Sebastian Stuker, Pierre Godard, Markus Muller, Lucas Ondel, Shruti Palaskar, Philip Arthur, Francesco Ciannella, Mingxing Du, Elin Larsen, Danny Merkx, Rachid Riad, Liming Wang, and Emmanuel Dupoux. Linguistic Unit Discovery from multi-modal inputs in unwritten languages: Summary of the "Speaking Rosetta" JSALT 2017 Workshop. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, April 2018.

[48] Odette Scharenborg, Francesco Ciannella, Shruti Palaskar, Alan Black, Florian Metze, Lucas Ondel, and Mark Hasegawa-Johnson. Building an asr system for a low-resource language through the adaptation of a high-resource language asr system: Preliminary results. In *Proceedings of ICNLSSP, Casablanca, Morocco*, 2017.

[49] Odette Scharenborg, Patrick Ebel, Francesco Ciannella, Mark Hasegawa-Johnson, and Najim Dehak. Building an asr system for mboshi using a cross-language definition of acoustic units approach. In *Proceedings of SLTU*, 2018.

[50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image classification. https://arxiv.org/abs/1409.1556 (2014). Accessed: 2017-09-14, 2017.

[51] Sebastian Stüker et al. Innovative technologies for under-resourced language documentation: The Bulb project. In *Proceedings of CCURL*, Portorož Slovenia, 2016.

[52] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[53] Alex Waibel Tanja Schultz. Experiments on cross-language acoustic modelling. In *Proceedings of Interspeech*, 2001.

[54] Tomoki Toda, Alan W. Black, and Keiichi Tokuda. Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis. In *Proceedings of SSW5, Pittsburgh, PA*, pages 31–36, 2004.

[55] Balakrishnan Varadarajan, Sanjeev Khudanpur, and Emmanuel Dupoux. Unsupervised learning of acoustic sub-word units. In *Proceedings of ACL on Human Language Technologies: Short Papers*, pages 165–168, 2008.

[56] K. Vesely, M. Karafi?t, F. Grezl, M. Janda, and E. Egorova. The language-independent bottleneck features. In *Proceedings of SLT*, 2012.

[57] Ngoc Thang Vu, Florian Metze, and Tanja Schultz. Multilingual bottleneck features and its application for under-resourced languages. In *Proc. 3rd Workshop on Spoken Language Technologies for Under-resourced Languages*, Cape Town; S. Africa, May 2012. MICA.

[58] Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. Sequence-to-sequence models can directly transcribe foreign speech. *arXiv preprint arXiv:1703.08581*, 2017.

[59] 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 *Proceedings of Interspeech*, pages 2132–2136, 2015.

[60] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4651–4659, 2016.

[61] Y. Zhang and J. R. Glass. Towards multi-speaker unsupervised speech pattern discovery. In *Proceeding of ICASSP*, pages 4366–4369, 2010.