UWSpeech: Speech to Speech Translation for Unwritten Languages

Chen Zhang\textsuperscript{1}, Xu Tan\textsuperscript{2}, Yi Ren\textsuperscript{1}, Tao Qin\textsuperscript{2}, Kejun Zhang\textsuperscript{1}, Tie-Yan Liu\textsuperscript{2}

\textsuperscript{1}Zhejiang University, China
\textsuperscript{2}Microsoft Research Asia

taoqin@microsoft.com, zhangkejun@zju.edu.cn, tyliu@microsoft.com

Abstract

Existing speech to speech translation systems heavily rely on the text of target language: they usually translate source language either to target text and then synthesize target speech from text, or directly to target speech with target text for auxiliary training. However, those methods cannot be applied to unwritten target languages, which have no written text or phoneme available. In this paper, we develop a translation system for unwritten languages, named as UWSpeech, which converts target unwritten speech into discrete tokens with a converter, and then translates source-language speech into target discrete tokens with a translator, and finally synthesizes target speech from target discrete tokens with an inverter. We propose a method called XL-VAE, which enhances vector quantized variational autoencoder (VQ-VAE) with cross-lingual (XL) speech recognition, to train the converter and inverter of UWSpeech jointly. Experiments on Fisher Spanish-English conversation translation dataset show that UWSpeech outperforms direct translation and VQ-VAE baseline by about 16 and 10 BLEU points respectively, which demonstrate the advantages and potentials of UWSpeech.

1 Introduction

Speech to speech translation (Lavie et al. 1997; Nakamura et al. 2006; Wahlster 2013; Jia et al. 2019) is important to help the understanding of cross-lingual spoken conversations and lectures, and has been used in scenarios such as international travel or conference. Existing speech to speech translation systems either rely on target text as a pivot (they first translate source speech into target text and then synthesize target speech given the translated text (Lavie et al. 1997; Nakamura et al. 2006; Wahlster 2013)), or directly translate source speech into target speech (Jia et al. 2019). In these translation systems, the text corresponding to the target speech is leveraged as either pivots or auxiliary training data (Jia et al. 2019); otherwise, the translation would not be possible or the translation accuracy would drop dramatically (Jia et al. 2019).

However, there are thousands of unwritten languages in the world (Scharenborg et al. 2020; Godard et al. 2017), which are purely spoken and have no written text. It is challenging to build speech translation systems for these unwritten languages without text as pivots or auxiliary training data. However, those methods cannot be applied to unwritten target languages, which have no written text or phoneme available. In this paper, we develop a translation system for unwritten languages, named as UWSpeech, which converts target unwritten speech into discrete tokens with a converter, and then translates source-language speech into target discrete tokens with a translator, and finally synthesizes target speech from target discrete tokens with an inverter. We propose a method called XL-VAE, which enhances vector quantized variational autoencoder (VQ-VAE) with cross-lingual (XL) speech recognition, to train the converter and inverter of UWSpeech jointly. Experiments on Fisher Spanish-English conversation translation dataset show that UWSpeech outperforms direct translation and VQ-VAE baseline by about 16 and 10 BLEU points respectively, which demonstrate the advantages and potentials of UWSpeech.

like in Jia et al. (2019). Continuous speech (which usually contains content, context, speaking style, etc.) is much more flexible to represent semantic meanings than discrete symbols (text) (van den Oord, Vinyals et al. 2017; Vigliocco et al. 2004), which makes the translation into speech harder than translation into text. Therefore, the key to ease the speech translation for unwritten languages is to reduce the flexible continuous space of speech into a more restricted discrete space.

A variety of previous works (Muthukumar and Black 2014; Chen et al. 2015; Wilkinson, Zhao, and Black 2016; Kamper, Livescu, and Goldwater 2017; Dunbar et al. 2017; Eloff et al. 2019; Tjandra et al. 2019; Duong et al. 2016; Salesky, Sperber, and Black 2019) have investigated the conversion between speech and their corresponding phonetic categories (discrete tokens) in an unsupervised manner, which mimics the way that human infants learn acoustic models in their mother tongue during their early years of life (Versteegh et al. 2016) (some of them only focus on a much easier task such as speech-to-text translation (Duong et al. 2016; Salesky, Sperber, and Black 2019)). Among these works, vector quantized variational autoencoder (VQ-VAE) (van den Oord, Vinyals et al. 2017; Dunbar et al. 2017; Tjandra et al. 2019; Chorowski et al. 2019; Liu et al. 2019; Tjandra, Sakti, and Nakamura 2019; Baevski, Schneider, and Auli 2019) has been widely adopted and shown advantages over other methods. However, VQ-VAE is still purely unsupervised and cannot ensure the quality of the learned discrete representations. Therefore, although VQ-VAE performs very well on relatively easier tasks like speech synthesis (Dunbar et al. 2019), it cannot achieve good accuracy on more complicated speech to speech translation where semantic representations of speech are important and more accurate phonetic representations are required. Few works tackle on speech to speech translation for unwritten languages (Tjandra, Sakti, and Nakamura 2019) since it is extremely challenging.

In this paper, we develop UWSpeech (UW is short for Un Written), a translation system for unwritten languages with three key components: 1) a converter that transforms unwritten target speech into discrete tokens, 2) a translator that translates source-language speech into target-language discrete tokens, and 3) an inverter that converts the translated discrete tokens back to unwritten target speech. As can be
seen, the discretization (transform speech into discrete tokens using converter) and reconstruction (synthesize speech from discrete tokens using inverter) steps in UWSpeech is important to ensure translation accuracy.

To this end, we propose XL-VAE, which improves the discretization and reconstruction capability based on VQ-VAE. Different from VQ-VAE that purely relies on unsupervised methods for discrete representation learning, XL-VAE leverages written languages with phonetic labels to improve the vector quantization (discrete representations learning) of unwritten languages through cross-lingual (XL) transfer. As human beings share similar vocal organs and pronunciations (Wind 1989), no matter which spoken languages they use, the phonetic representations learned in one language can more or less (depending on the language similarity) help the learning of phonetic representations in another language (Kuhl et al. 2008). Therefore, XL-VAE can benefit from other written languages and outperform purely unsupervised VQ-VAE on discretizing speech into discrete tokens and synthesizing speech from discrete tokens, and thus enable UWSpeech to achieve better translation accuracy.

Our contributions can be summarized as follows:

- We develop UWSpeech, a speech to speech translation system for unwritten languages, and design a novel XL-VAE to train the converter and inverter in UWSpeech jointly for discrete speech representations.
- We conduct experiments on Fisher Spanish-English speech conversation dataset, assuming the target language is unwritten. Experiment results show that UWSpeech equipped with XL-VAE achieves 16 and 10 BLEU points improvements over direct translation and VQ-VAE baseline respectively, which demonstrates the advantages and potentials of UWSpeech on speech to speech translation for unwritten target languages. \(^1\)
- We further apply UWSpeech to text to speech translation and speech to text translation for unwritten languages. The improvements over direct translation and VQ-VAE baseline demonstrate the general applicability of UWSpeech beyond speech to speech translation.

\section{Background}

\subsection{A Taxonomy of Speech Translation and Our Focused Setting}

Based on the successes of text to text translation (Bahdanau, Cho, and Bengio 2014; Luong, Pham, and Manning 2015; Vaswani et al. 2017), speech translation (Bérard et al. 2016; Weiss et al. 2017; Jia et al. 2019) has been developed to handle speech as translation input and/or output. Previous works on speech translations has evolved from cascaded models (Ney 1999; Matusov, Kanthak, and Ney 2005; Lavie et al. 1997; Nakamura et al. 2006; Wahlster 2013) to end-to-end models (Bérard et al. 2016; Weiss et al. 2017; Vila et al. 2018; Sperber et al. 2019; Jia et al. 2019), where the text corresponding to speech is leveraged as auxiliary training (Jia et al. 2019) for better accuracy. Depending on the speech is in the source or/and target side, speech translation can be divided into three categories: speech to text translation, text to speech translation and speech to speech translation. In this paper, we focus on the most difficult setting: speech to speech translation for unwritten languages. In this way, we can not leverage any source or target text in auxiliary tasks like in Jia et al. (2019).

Furthermore, we also extend UWSpeech for text to speech translation with unwritten target languages and speech to text translation with unwritten source languages to demonstrate the generalization ability of our method. Besides, our method can also be applied to the written target languages whose text or phonetic transcripts are not available in the training data.

\subsection{Discrete Speech Representations}

Learning discrete representations of speech has long been studied for better speech understanding and modeling. Previous works on discrete speech representations include k-means clustering (Kamper, Livescu, and Goldwater 2017; Dunbar et al. 2017), Gaussian mixture model clustering (Chen et al. 2015), tree-based clustering (Muthukumar and Black 2014), binary with straight-through estimation (Elloff et al. 2019), categorical VAE (Elloff et al. 2019) and the more advanced vector quantized VAE (VQ-VAE) (van den Oord, Vinyals et al. 2017; Dunbar et al. 2019; Tjandra et al. 2019; Chorowski et al. 2019; Liu et al. 2019; Tjandra, Sakti, and Nakamura 2019; Baevski, Schneider, and Auli 2019). VQ-VAE has been widely used to cluster/quantize the representations of speech and discretize into codebook sequence, and has achieved good results on some tasks such as subword units discovery from speech or text to speech synthesis (Dunbar et al. 2019). However, VQ-VAE is a purely unsupervised clustering method for discrete speech representations, which limits its effectiveness on harder tasks like speech translation. In this paper, we improve VQ-VAE with cross-lingual (XL) speech recognition and propose XL-VAE to achieve better discrete speech representations.

\section{UWSpeech}

In this section, we introduce the design of our proposed UWSpeech: a speech to speech translation system for unwritten target languages with the help of cross-lingual vector quantized variational autoencoder (XL-VAE). We first describe the overall pipeline of UWSpeech, and then introduce the detailed design of XL-VAE.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{pipeline.png}
\caption{The training and inference pipeline of UWSpeech.}
\end{figure}

\footnote{Speech samples and experimental details can be found in \url{https://speechresearch.github.io/uwspeech/}}
3.1 Pipeline Overview

For speech to speech translation where the target language is unwritten, UWSpeech consists of three components as shown in Figure 1: 1) a converter to transform the target-language speech into discrete tokens; 2) a translator to translate the source speech into target discrete tokens; 3) an inverter to convert the target discrete tokens back to target speech. We introduce each component in the following subsections.

**Translator** Denote the training corpus as \( \{(x, y) \in (\mathcal{X}, \mathcal{Y})\} \), where \( x \) and \( y \) are the source and target speech sequence. According to the pipeline of UWSpeech, we convert the target unwritten speech sequence \( y \in \mathcal{Y} \) into discrete token sequence \( z \in \mathcal{Z} \) to form a triple corpus \( (\mathcal{X}, \mathcal{Z}, \mathcal{Y}) \). We train a machine translator \( \theta_{\text{trans}} \) by minimizing the negative log-likelihood loss

\[
L_{\text{trans}} = - \sum_{(x, z, y) \in (\mathcal{X}, \mathcal{Z}, \mathcal{Y})} \log P(z|x; \theta_{\text{trans}}),
\]

where \( \theta_{\text{trans}} \) can be implemented as a standard encoder-attention-decoder (Vaswani et al. 2017) based model with several convolution layers in the encoder to handle speech input, and will be described in the experiment setting.

**Converter and Inverter** The converter and inverter transform the speech sequence \( y \) into discrete token sequence \( z \) and transform \( z \) back to speech sequence \( y \) respectively, and follow the form of autoencoder where the converter acts like the encoder and the inverter acts like the decoder. Inspired by VQ-VAE, we propose a novel XL-VAE to better train the converter and inverter for speech translation.

3.2 XL-VAE

XL-VAE first encodes the speech sequence into hidden representations to extract discrete tokens with a converter, and reconstructs the original speech sequence given the representations of discrete tokens with an inverter. Different from VQ-VAE (van den Oord, Vinyals et al. 2017), XL-VAE extracts discrete representations not by unsupervised vector clustering, but by speech/phoneme recognition, where the recognition capability is transferred from other popular written languages. We train the phoneme recognition on written languages with speech and phoneme pairs based on the converter. We illustrate XL-VAE in Figure 2 and formulate each module in XL-VAE as follows.

**Converter** The converter of XL-VAE \( \theta_{\text{conv}} \) takes speech sequence \( y \) as input and generate continuous hidden representations \( z' \):

\[
z' = g(y; \theta_{\text{conv}}).
\]

\( z' \) is further converted into discrete latent variables \( z \) through nearest neighbour search based on dot-product:

\[
q(z|y) = \begin{cases} 1 & \text{for } \hat{z} = \arg \max_i (\hat{z} + e_i), \\ 0 & \text{otherwise} \end{cases},
\]

where \( q(z|y) \) denotes the categorical distribution of the discrete variable \( z \). \( e \in \mathbb{R}^{K \times D} \) denotes the embedding space of the discrete tokens, \( K \) denotes the number of discrete tokens and \( D \) denotes the size of each embedding vector \( e_i \) for \( i \in \{1, 2, ..., K\} \).

As shown in Figure 2, the converter takes speech (mel-spectrogram) sequence as input and uses several convolution layers with strides to reduce the length of speech sequence by \( 1/c \). It then stacks \( N \) Transformer blocks (Vaswani et al. 2017), where each block contains a self-attention layer and a feed-forward layer with a layer-normalization and a residual connection on top of each layer. For a speech sequence with length of \( l \), the generated discrete tokens \( z \) has length of \( l/c \).

![Figure 2: The model structure of XL-VAE.](image)

**Inverter** The inverter of XL-VAE \( \theta_{\text{inv}} \) takes discrete tokens \( z \) as input and convert \( z \) into \( e_z \) with discrete token look-up table \( e \) (the same \( e \) as used in the converter). Then \( e_z \) is used to reconstruct the original speech sequence \( y \):

\[
L_{\text{inv}} = \sum_{y \in \mathcal{Y}} (y - f(e_z; \theta_{\text{inv}}))^2.
\]

As shown in Figure 2, the inverter leverages several transposed convolution layers (Dumoulin and Visin 2016) to increase the length of \( e_z \) by \( c \times \) (opposed to the \( 1/c \times \) in the converter), to match the length of the original mel-spectrogram sequence. It then stacks \( N \) Transformer blocks (Vaswani et al. 2017) as used in the converter. The then transformed through a softmax function to get the probability of each phoneme category (which is described in the later part of this subsection).
 reverberant speech and phoneme sequence pairs in the discrete representations, as shown in Figure 2. Given the supervised quantization in VQ-VAE, XL-VAE introduces an inverter to further convert the mel-spectrogram into an audio waveform.

Cross-Lingual (XL) Speech Recognition Instead of unsupervised quantization in VQ-VAE, XL-VAE introduces speech recognition in other written languages to help learn the discrete representations, as shown in Figure 2. Given the speech and phoneme sequence pairs \((y', t') \in (Y', T')\) of written languages, we use the converter \(\theta_{\text{conv}}\) to transform speech \(y'\) into \(\hat{z}'\), and then multiply \(\hat{z}'\) with the discrete token embedding matrix \(e\) (in denoted in Equation 3) and get the probability distribution \(P_{\text{inv}}\) over \(K\) phoneme categories with a softmax operation, where \(K\) is size of phoneme vocabulary in the written languages, and also the number of discrete tokens in \(e\), which is similar with Li et al. (2020). We train the phoneme recognition with connectionist temporal classification (CTC) loss (Graves et al. 2006). The formulation of the cross-lingual speech recognition is as follows:

\[
\hat{z}' = f(y'; \theta_{\text{conv}}), \quad P_{\text{inv}}(r) = \prod_{i=1}^{\mid r \mid} \text{softmax}(\hat{z}' + e)_{r_i},
\]

\[
L_{\text{ctc}} = - \sum_{{(x', t') \in (Y', T')} \in \phi(t')} \sum_{s \in \phi(t')} \log P_{\text{inv}}(r = s),
\]

where \(\phi(t')\) denotes the set of valid CTC paths for phoneme sequence \(t'\), \(P_{\text{inv}}(r = s)\) denotes the probability of the CTC path \(s\), softmax\()_{r_i}\) denotes the probability of observing label \(r_i\) under the softmax function and \(\mid r \mid\) denotes the length of sequence \(r\). The loss function \(L_{\text{ctc}}\) aims to minimize the negative log-likelihood of all the valid CTC paths in the training set. For more details of CTC, you can refer to Graves et al. (2006), which is not the focus of this work.

Discrete Representation We choose international phonetic alphabet (IPA) (Association, Staff et al. 1999) as the phoneme set of the written languages. In this way, the discrete token embeddings \(e \in \mathcal{R}^{K \times D}\) are exactly the embeddings of IPA where \(K\) is the size of IPA set and \(D\) is the dimension of the embedding vector. The unwritten speech is converted into discrete tokens which fall into the IPA set of written languages. The discrete tokens \(z\) as well as the corresponding embedding vectors in \(e\) are taken as the discrete representations of speech \(y\).

Loss Function of XL-VAE Putting Equation 2, 3, 4 and 5 together, we have the loss function of XL-VAE:

\[
L_{\text{xl-vae}} = L_{\text{inv}} + \lambda L_{\text{xl}},
\]

where \(\lambda\) is a hyperparameter to trade-off the two loss terms.

3.3 Training and Inference
Finally, we describe the training and inference procedure of UWSpeech according to the formulations in the previous two subsections. The detailed procedure is shown in Algorithm 1.

Algorithm 1 UWSpeech Training and Inference

Training:
Input: Speech to speech translation corpus \((X, Y)\) where \(Y\) represents target unwritten speech. Paired speech and phoneme corpus \((Y', T')\) in written languages where \(T'\) uses IPA as the phoneme set.
Step 1: Train the XL-VAE model with corpus \(Y\) and \((Y', T')\) using loss in Equation 6 to obtain the converter \(\theta_{\text{conv}}\), inverter \(\theta_{\text{inv}}\) and discrete token look-up table \(e\).
Step 2: Convert the unwritten speech corpus \(Y\) into discrete sequence corpus \(Z\) following Equation 2 and 3. Train the machine translator \(\theta_{\text{trans}}\) with corpus \((X, Z)\) using loss in Equation 1.

Inference:
Input: Source speech corpus \(X\), translator \(\theta_{\text{trans}}\), discrete token look-up table \(e\) and inverter \(\theta_{\text{inv}}\).
Step 1: For each speech sequence \(x \in X\), generate target discrete tokens: \(z \sim P(z|x; \theta_{\text{trans}})\).
Step 2: Convert \(z\) into \(e\), through discrete token look-up table \(e\), and synthesize target speech: \(y = f(e, z; \theta_{\text{inv}})\).

4 Experiments and Results
In this section, we first introduce the experimental setup and then report the results of UWSpeech for speech to speech translation. We further conduct some analyses of UWSpeech. Finally, we also apply UWSpeech to text to speech translation and speech to text translation settings.

4.1 Experimental Setup

Datasets Following the common practice in low-resource and unsupervised speech and translation works (Lample et al. 2018; Song et al. 2019; Ren et al. 2019), we conduct experiments on popular written languages but remove the text of target speech to simulate unwritten languages. We choose Fisher Spanish–English dataset (Post et al. 2013) for translation. Considering 1) translation to unwritten languages is difficult and 2) the most useful translation scenarios for unwritten languages are daily communication, travel translation, etc., where high-frequency and simple words/sentences are usually used, we choose some common sentences from the original full test set to form our test set (denoted as common test set). But we still show the results on the full test set of the main experiments setting in Table 1 and Table 2 for reference. For the written languages used in XL-VAE, we choose French, German and Chinese with speech data and corresponding phoneme sequence. Both the German and French datasets are from Common Voice\(^3\), where the German corpus contains about 280K training examples (325 hours) with 5007 different speakers and the French corpus contains 150K training examples (173 hours)

\(^3\)https://voice.mozilla.org/
with 3005 different speakers. For the Chinese dataset, we use AIShell (Bu et al. 2017) which contains about 140K training examples (178 hours) with 400 different speakers.

Model Configuration We choose Transformer (Vaswani et al. 2017) as the basic model structure for the converter, inverter and translator, since it achieves good results on machine translation, speech recognition and speech synthesis tasks.

Training Details We first train the converter, inverter and discrete token embeddings in XL-VAE. We up-sample the speech data of each written language (German, French, Chinese) to the same amount, and then up-sample the speech data of unwritten language (English or Spanish) to match the total amount of written languages. We ensure there are an equal amount of data in written and unwritten languages in each mini-batch. We choose the λ in Equation 6 according to the validation performance and set λ to 0.01. The batch size is set to 25K frames for each GPU and the XL-VAE training takes 200K steps on 4 Tesla V100 GPUs.

After the training of XL-VAE, the phoneme error rates (PER) of three written languages (German, French and Chinese) on the development set are 16%, 21% and 12% respectively. We convert the target unwritten speech into the discrete token sequence and keep the output discrete token sequence as it is, without removing any special or repeated tokens. We use the discrete token sequence generated by XL-VAE to train translator, with a batch size of 16K frames on each GPU and 100K training steps on 4 Tesla V100 GPUs.

Our code is implemented based on tensor2tensor library (Vaswani et al. 2018) 1.

Inference and Evaluation During inference, we use the translator to generate discrete token sequences from source speech with beam search. We set beam size to 4 and the length penalty to 1.0. We then directly use the inverter to transform the discrete token sequence back to target speech.

To evaluate the accuracy of the speech translation, following the practice in Jia et al. (2019), we pre-train an automatic speech recognition model (which can achieve 85.62 BLEU points on our test set and is comparable with Jia et al. (2019)) to generate the corresponding text of the translated speech, and then calculate the BLEU score (Papineni et al. 2002) between the generated text and the reference text. We report BLEU score using case insensitive BLEU with moses tokenizer and multi-bleu.perl 2. Due to the Fisher corpus has 4 English references in the test set, we report 4-reference BLEU score for Spanish to English setting, and still report single-reference BLEU score for English to Spanish setting.

4https://github.com/tensorflow/tensor2tensor

4.2 Results
In this subsection, we report the experiment results of UWSpeech. We compare UWSpeech mainly with two baselines: 1) Direct Translation (denoted as Direct), which directly translates the source speech into target speech in an encoder-attention-decoder model without any text as auxiliary training data or pivots. 2) Discretization with VQ-VAE (denoted as VQ-VAE), which follows the translation pipeline in UWSpeech but replaces XL-VAE with original VQ-VAE for speech discretization.

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>VQ-VAE</th>
<th>UWSpeech</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common (Es→En)</td>
<td>1.45</td>
<td>7.17</td>
<td>17.33</td>
</tr>
<tr>
<td>Full (Es→En)</td>
<td>0.8</td>
<td>3.42</td>
<td>9.35</td>
</tr>
<tr>
<td>Common (En→Es)</td>
<td>0.80</td>
<td>3.12</td>
<td>11.13</td>
</tr>
<tr>
<td>Full (En→Es)</td>
<td>0.62</td>
<td>1.45</td>
<td>8.27</td>
</tr>
</tbody>
</table>

Table 1: The BLEU scores of speech to speech translation on two translation directions, where Common means common test set and Full means full test set.

The speech to speech translation results on Spanish to English are shown in Table 1. As can be seen, Direct achieves a very low BLEU score, which is consistent with the findings in Jia et al. (2019) and demonstrates the difficulty of direct speech to speech translation. VQ-VAE achieves slightly better BLEU score than Direct, but still with poor accuracy, which demonstrates the limitations of the purely unsupervised method for speech discretization when handling speech translation. On the common test set as we described in Section 4.1, UWSpeech achieves 17.33 BLEU points, about 10 points higher than VQ-VAE and 16 points higher than Direct. UWSpeech also shows a huge gain on the full test set. We also find that the inverter in XL-VAE can get a lower reconstruction loss than VQ-VAE on the validation set, demonstrating that the discrete tokens extracted by XL-VAE can not only help the discrete token translation in translator but can also benefit the speech reconstruction in inverter, which together contributes to the better accuracy in speech translation. The above results demonstrate the advantages of XL-VAE in leveraging cross-lingual speech recognition for speech discretization and the effectiveness of UWSpeech for unwritten speech translation.

The experiment results on English to Spanish translation are also shown in Table 1. Similar to the results on Spanish to English translation, Direct achieves a very low BLEU score and UWSpeech achieves about 8 points higher than VQ-VAE on the common test set and 7 points higher on the full test set, demonstrating the effectiveness of UWSpeech.

4.3 Method Analyses
In this subsection, we conduct some experimental analyses on the proposed UWSpeech. For simplicity, we only show the results on the common test set we described in Section 4.1.

UWSpeech with Multi-task Training Jia et al. (2019) proposes a direct speech to speech translation model, which
improves translation accuracy through multi-task training (source speech to source text (automatic speech recognition), and source speech to target text (speech to text translation)). Originally, due to lack of text in both source and target languages, speech to speech translation for unwritten languages could not take advantage of the multi-task training mechanism. However, our proposed XL-V AE can discretize the speech into discrete tokens, which can be regarded as text for multi-task training. Therefore, we study how UWSpeech performs when combining with multi-task training.

We combine UWSpeech with multi-task training in two ways:

- SL ASR (Source Language ASR): Training a model that has a shared speech encoder and two decoders: one is for speech recognition on source unwritten languages (source speech to the corresponding discrete tokens), and the other is for speech translation on source unwritten languages (source speech to the discrete tokens in the target language). Both of the discrete tokens corresponding to the source and target unwritten languages are generated by XL-VAE. In this way, we leverage automatic speech recognition of source unwritten language (discrete token sequences as target) as auxiliary loss in our Translator.

- WL ASR (Written Languages ASR): Training a model that has a shared speech encoder and two decoders: one is for phone-level automatic speech recognition on auxiliary written languages (e.g., German, French, and Chinese in this paper), and the other is for speech translation on unwritten languages (e.g., translate Spanish speech to English speech directly) at the same time, hoping that ASR can help the speech encoder training better.

As we can see in Table 2, the SL ASR setting can only improve slightly from 17.33 to 17.41, which also demonstrates the discretization of source speech is not so necessary. The BLEU score of the WL ASR setting is very low (2.36), which indicates that the Direct Translation model cannot make full use of the written languages, while XL-VAE can do this well.

<table>
<thead>
<tr>
<th>Method</th>
<th>UWSpeech</th>
<th>SL ASR</th>
<th>WL ASR</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>17.33</td>
<td>17.41</td>
<td>2.36</td>
</tr>
</tbody>
</table>

Table 2: The BLEU scores of Spanish to English speech to speech translation, combines with multi-task training in different ways.

Analyses of Written Languages in XL-VAE We study the influence of written languages in XL-VAE on the translation accuracy, mainly from two perspectives: 1) the data amount of the written languages, and 2) the similarity between the written and unwritten languages. To this end, we design several different experimental settings for this study, as shown in Table 3.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Configuration</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>De (80h)</td>
<td>10.58</td>
</tr>
<tr>
<td>#2</td>
<td>De (160h)</td>
<td>12.12</td>
</tr>
<tr>
<td>#3</td>
<td>De (320h)</td>
<td>15.20</td>
</tr>
<tr>
<td>#4</td>
<td>De (320h) + Fr (160h) + Zh (160h)</td>
<td>17.33</td>
</tr>
<tr>
<td>#5</td>
<td>Fr (160h)</td>
<td>11.79</td>
</tr>
<tr>
<td>#6</td>
<td>Zh (160h)</td>
<td>9.38</td>
</tr>
</tbody>
</table>

Table 3: The BLEU scores of Spanish to English translation with different written languages as well as different data amounts for XL-VAE. We denote German as De, French as Fr and Chinese as Zh.

Varying Embedding Size $D$ and Down-Sampling Ratio $c$ in XL-VAE We further evaluate how the discrete token embedding size $D$ and the speech down-sampling ratio $c$ in XL-VAE influence the translation accuracy. We set $c = 4$ when varying $D$ and set $D = 256$ when varying $c$ according to preliminary experiments. As shown in Table 4, discrete token embedding size $D = 256$ performs better and down-sampling ratio $c = 4$ performs better.

<table>
<thead>
<tr>
<th>Embedding Size $D$</th>
<th>64</th>
<th>128</th>
<th>256</th>
<th>512</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>13.85</td>
<td>15.20</td>
<td>17.33</td>
<td>17.13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Down-Sampling Ratio $c$</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>10.05</td>
<td>13.27</td>
<td>17.33</td>
<td>16.85</td>
</tr>
</tbody>
</table>

Table 4: The BLEU scores of Spanish to English translation with different discrete token embedding sizes and down-sampling ratios.

The Advantage of Training Converter and Inverter Jointly To study the benefits of joint training the conversion and language similarity representation instead of acoustic conditions considering the good robustness of the ASR model.
In this paper, we developed UWSpeech, a speech to speech translation system for unwritten target languages, and designed XL-VAE, an enhanced version of VQ-VAE based on cross-lingual speech recognition, to jointly train the converter and inverter to discretize and reconstruct the unwritten speech in UWSpeech. Experiments on Fisher Spanish-English dataset demonstrate that UWSpeech equipped with XL-VAE achieves significant improvements in translation accuracy over the direct translation and VQ-VAE baseline.

In the future, we will enhance XL-VAE with domain adversarial training to better transfer the speech recognition ability from written languages to unwritten languages. We will test UWSpeech on more complicated sentences and language pairs. Furthermore, going beyond the proof-of-concept experiments in this work (we assumed English or Spanish is unwritten), we will apply UWSpeech on truly unwritten languages for speech to speech translation.
Acknowledgments

This work was supported by the Key Project of Natural Science Foundation of Zhejiang Province (No. LZ19F020002). This work was also partially funded by Microsoft Research Asia. Thanks are due to Shen Zhou for bringing strength during tough times.

References


