UWSpeech: Speech to Speech Translation for Unwritten Languages

Chen Zhang[1], Xu Tan[2], Yi Ren[1], Tao Qin[2], Kejun Zhang[1], Tie-Yan Liu[2]
[1]Zhejiang University, China
[2]Microsoft Research Asia
zc99@zju.edu.cn, xuta@microsoft.com, rayeren@zju.edu.cn,
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 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. 2019; 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 UnWritten), 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.

2 Background

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.

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), binarization with straight-through estimation (Eloff et al. 2019), categorical VAE (Eloff 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.

3 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.

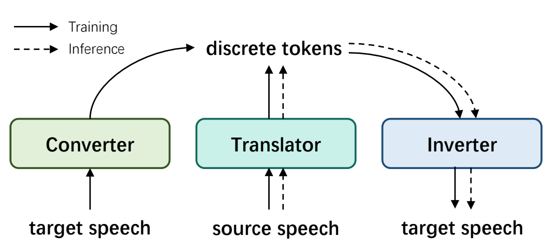

Figure 1: The training and inference pipeline of UWSpeech.

[1] Speech samples and experimental details can be found in 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 θ_{trans} by minimizing the negative log-likelihood loss

$$\mathcal{L}_{\text{trans}} = -\sum_{(x,z)\in(\mathcal{X},\mathcal{Z})} \log P(z|x;\theta_{\text{trans}}), \tag{1}$$

where θ_{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 θ_{conv} takes speech sequence y as input and generate continuous hidden representations $\hat{z}$:

$$\hat{z} = f(y;\theta_{\text{conv}}). \tag{2}$$

$\hat{z}$ is further converted into discrete latent variables z through nearest neighbour search based on dot-product[2]:

$$q(z=k|y) = \begin{cases} 1 & \text{for } k = \arg\max_i \ (\hat{z} * e_i) \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

where $q(z|y)$ denotes the categorical distribution of the discrete variable z. $e \in \mathcal{R}^{K\times D}$ 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.

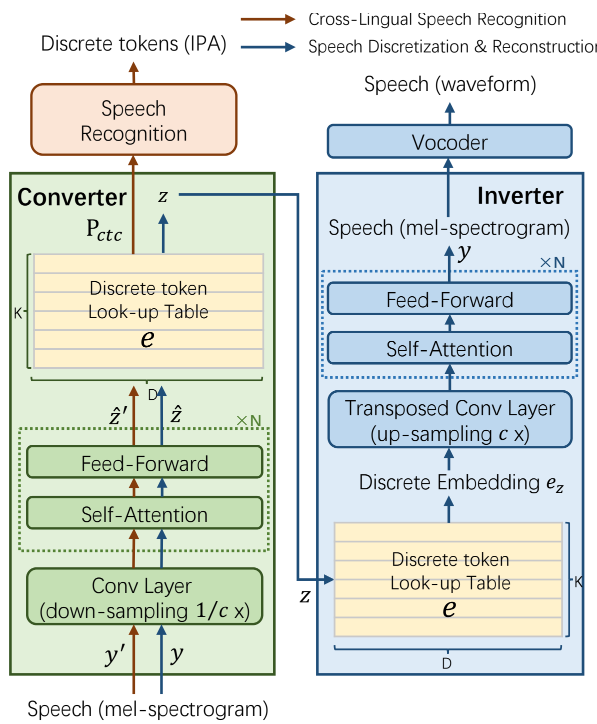

Figure 2: The model structure of XL-VAE.

Inverter The inverter of XL-VAE θ_{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:

$$\mathcal{L}_{\text{inv}} = \sum_{y\in\mathcal{Y}} (y - f(e_z;\theta_{\text{inv}}))^2. \tag{4}$$

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

[2] We use dot-product here instead of Euclidean distance in VQ-VAE, in order to be consistent with the speech recognition where the hidden representations $\hat{z}'$ are multiplied with the matrix e and then transformed through a softmax function to get the probability of each phoneme category (which is described in the later part of this subsection).

inverter reconstructs the speech sequence in parallel. Therefore, different from the conventional self-attention in Transformer decoder which cannot see the information in the future positions, the self-attention in the inverter can see the information in all positions, just like the converter. A vocoder (Griffin and Lim 1984; Oord et al. 2016) is leveraged 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 $(y', t') \in (\mathcal{Y}', \mathcal{T}')$ of written languages, we use the converter θ_{conv} to transform speech y' into $\hat{z}'$, and then multiply $\hat{z}'$ with the discrete token embedding matrix e (e is denoted in Equation 3) and get the probability distribution P_{ctc} 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:

$$\hat{z}' = f(y'; \theta_{\text{conv}}), \quad P_{\text{ctc}}(r) = \prod_{i=1}^{|r|} \text{softmax}(\hat{z}' * e)_{r_i},$$
$$\mathcal{L}_{\text{xl}} = - \sum_{(x',t') \in (\mathcal{Y}', \mathcal{T}')} \sum_{s \in \phi(t')} \log P_{\text{ctc}}(r = s), \tag{5}$$

where $\phi(t')$ denotes the set of valid CTC paths for phoneme sequence t', $P_{\text{ctc}}(r = s)$ denotes the probability of the CTC path s, $\text{softmax}(\cdot)_{r_i}$ denotes the probability of observing label r_i under the softmax function and $|r|$ denotes the length of sequence r. The loss function $\mathcal{L}_{\text{xl}}$ 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:

$$\mathcal{L}_{\text{xl-vae}} = \mathcal{L}_{\text{inv}} + \lambda \mathcal{L}_{\text{xl}}, \tag{6}$$

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

Inference:
Input: Source speech corpus $\mathcal{X}$, translator θ_{trans}, discrete token look-up table e and inverter θ_{inv}.
Step 1: For each speech sequence $x \in \mathcal{X}$, generate target discrete tokens: $z \sim P(z|x; \theta_{\text{trans}})$.
Step 2: Convert z into e_z 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 **??** 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)[4].

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[5] and multi-bleu.perl[6]. 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.

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

[5]https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

[6]https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

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.

Method	*Direct*	*VQ-VAE*	*UWSpeech*
Common (Es→En)	1.45	7.17	17.33
Full (Es→En)	0.8	3.42	9.35
Common (En→Es)	0.80	3.12	11.13
Full (En→Es)	0.62	1.45	8.27

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-VAE 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 to 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.

Method	*UWSpeech*	*SL ASR*	*WL ASR*
BLEU	17.33	17.41	2.36

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[7].

[7]Someone may wonder the acoustic conditions of the speech in different written languages may influence the comparison. We listened and compared the acoustic conditions in their speech data and only found little difference. Therefore, we can focus more on the data amount and language similarity instead of acoustic conditions considering the good robustness of the ASR model.

From setting #1, #2 and #3, it can be seen that increasing the data amount of written language (German) can improve the speech translation accuracy. Comparing setting #4 with #3, we can find that further adding other languages (French and Chinese) to increase the total data amount can also improve translation accuracy. Comparing setting #2, #5 and #6, we can find that German helps more on the discretization of English than French, and both German and French help more than Chinese, which is consistent with the language similarity. According to the language family (Lewis, Simons, and Fennig 2009), German and English belong to the same Germanic branch in the Indo-European family, while French and English belong to the same Indo-European family although not in the same branch. Chinese and English belong to different families and are far apart from each other. Even using distant Chinese as written language, our method still achieves higher accuracy than VQ-VAE (9.38 vs 7.17).

Setting	Configuration	BLEU
#1	De (80h)	10.58
#2	De (160h)	12.12
#3	De (320h)	15.20
#4	De (320h) + Fr (160h) + Zh (160h)	17.33
#5	Fr (160h)	11.79
#6	Zh (160h)	9.38

Table 3: The BLEU scores of Spanish to English speech to speech 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.

Embedding Size D	64	128	256	512
BLEU	13.85	15.20	17.33	17.13
Down-Sampling Ratio c	1	2	4	8
BLEU	10.05	13.27	17.33	16.85

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 Inventer Jointly To study the benefits of joint training the converter and inverter in XL-VAE, we separately train the converter by speech recognition on written languages and the inverter by reconstructing speech from discrete tokens. Separate training achieves 13.51 BLEU points on Spanish to English translation, which is much lower than joint training in XL-VAE (17.33), demonstrating the effectiveness of joint training in XL-VAE. Also, the setting points out that even if we pre-train VQ-VAE with the same unwritten language data, it underperforms our UWSpeech.

Discretization of Source Speech To study the translation accuracy if we also discretize the source speech into discrete tokens at the same time, we conduct experiments on Spanish to English translation direction and achieve 17.45 BLEU points, which is just slightly better than only discretizing target speech (17.33). The results demonstrate that direct translation from source speech is not as difficult as direct translation into target speech.

Case Analyses We further analyze some translation cases by our UWSpeech system and the baseline methods on Spanish to English translation. As shown in Table 5, we list the source (Spanish) and target (English) reference text corresponding to the speech, and convert the translated English speech into text with the pre-trained automatic speech recognition model as used in the evaluation. For the first case, both *Direct Translation* and *VQ-VAE* miss the meaning of "what she said" while *UWSpeech* can translate the meaning. For the second case, only *UWSpeech* can translate the meaning of "How's it going, where are you from?" correctly. We also show the translated discrete token sequence (IPA) by the translator (denoted as *IPA (UWSpeech)*) as well as the discrete token sequence extracted from the target speech (denoted as *IPA (Target)*) in Table 5. It can be seen that the IPA translated by UWSpeech is close to the target IPA, and both are close to the pronunciation of English speech, which demonstrates the good accuracy of the IPA extracted by XL-VAE and translated by the translator. We attach the corresponding speech and more cases at https://speechresearch.github.io/uwspeech.

<table>
<tr><td>Spanish (Source)</td><td>Yo no entendí lo que ella dijo.</td></tr>
<tr><td>English (Target)</td><td>I didn't understand what she said.</td></tr>
<tr><td>Direct Translation</td><td>I don't know.</td></tr>
<tr><td>VQ-VAE</td><td>I didn't understand.</td></tr>
<tr><td>UWSpeech</td><td>I didn't understand what she say.</td></tr>
<tr><td>IPA (Target)</td><td>ai ai n | d ɪ g n n z ɛ | ɐ n ɤ s t t ə l ɐ n n t t | v ɔ t | t iː s | ɛ ɛː n</td></tr>
<tr><td>IPA (UWSpeech)</td><td>ai ai | d e n n n | a n n v yː s s t e n n n t | v ɔ ɐ t | t d iː | ɛ ɛ l</td></tr>
<tr><td>Spanish (Source)</td><td>Qué tal, ¿de dónde eres?</td></tr>
<tr><td>English (Target)</td><td>How's it going, where are you from?</td></tr>
<tr><td>Direct Translation</td><td>Had a price.</td></tr>
<tr><td>VQ-VAE</td><td>Like are you there are you from?</td></tr>
<tr><td>UWSpeech</td><td>How are you, where are you from?</td></tr>
<tr><td>IPA (Target)</td><td>h h aː s | b ɪ t t | ɔy n n | ɔ | ʁ | j | v aː ʁ m</td></tr>
<tr><td>IPA (UWSpeech)</td><td>h h au | ɐ | j øː | ɐ | j e | v aː ʁ m</td></tr>
</table>

Table 5: Some translation cases in Spanish to English speech to speech translation.

of our UWSpeech, these results also demonstrate that it is not that necessary to discretize the source speech in speech translation, which is consistent with our findings in Section 4.3, and is also consistent with the results in Weiss et al. (2017) where even leveraging the ground-truth text corresponding the source speech can only achieve a BLEU gain less than 2 points.

Method	*Direct Translation*	*VQ-VAE*	*UWSpeech*
Text to Speech	5.47	8.02	22.03
Speech to Text	33.87	29.98	34.05

Table 6: The BLEU scores of the text to speech and speech to text setting on Spanish to English translation, where English and Spanish is taken as the unwritten language in the text to speech setting and speech to text setting respectively.

4.4 Extension of UWSpeech

Although UWSpeech is designed for speech to speech translation, it can also be applied to other two speech translation settings for unwritten languages: text to speech translation and speech to text translation. We conduct experiments on these two settings on Spanish to English translation to verify the broad applicability of UWSpeech for unwritten speech translation, and show the results on the common test set we described in Section 4.1 in Table 6.

In the text to speech setting, *Direct Translation* still achieves very poor translation accuracy and *UWSpeech* achieves about 14 BLEU points improvements over *VQ-VAE* baseline, demonstrating the effectiveness of UWSpeech on text to speech translation for unwritten languages.

In the speech to text setting, *UWSpeech* achieves much higher accuracy than *VQ-VAE* and slightly better accuracy than *Direct Translation*. While verifying the effectiveness

5 Conclusion

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

Association, I. P.; Staff, I. P. A.; et al. 1999. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

Baevski, A.; Schneider, S.; and Auli, M. 2019. vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations. *arXiv preprint arXiv:1910.05453* .

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* .

Bérard, A.; Pietquin, O.; Servan, C.; and Besacier, L. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *arXiv preprint arXiv:1612.01744* .

Bu, H.; Du, J.; Na, X.; Wu, B.; and Zheng, H. 2017. AIShell-1: An Open-Source Mandarin Speech Corpus and A Speech Recognition Baseline. In *Oriental COCOSDA 2017*, Submitted.

Chen, H.; Leung, C.-C.; Xie, L.; Ma, B.; and Li, H. 2015. Parallel inference of Dirichlet process Gaussian mixture models for unsupervised acoustic modeling: A feasibility study. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Chorowski, J.; Weiss, R. J.; Bengio, S.; and van den Oord, A. 2019. Unsupervised speech representation learning using wavenet autoencoders. *IEEE/ACM transactions on audio, speech, and language processing* 27(12): 2041–2053.

Dumoulin, V.; and Visin, F. 2016. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285* .

Dunbar, E.; Algayres, R.; Karadayi, J.; Bernard, M.; Benjumea, J.; Cao, X.-N.; Miskic, L.; Dugrain, C.; Ondel, L.; Black, A. W.; and et al. 2019. The Zero Resource Speech Challenge 2019: TTS Without T. *Interspeech 2019* doi: 10.21437/interspeech.2019-2904. URL http://dx.doi.org/10.21437/interspeech.2019-2904.

Dunbar, E.; Cao, X. N.; Benjumea, J.; Karadayi, J.; Bernard, M.; Besacier, L.; Anguera, X.; and Dupoux, E. 2017. The zero resource speech challenge 2017. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 323–330. IEEE.

Duong, L.; Anastasopoulos, A.; Chiang, D.; Bird, S.; and Cohn, T. 2016. An attentional model for speech translation without transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 949–959.

Eloff, R.; Nortje, A.; van Niekerk, B.; Govender, A.; Nortje, L.; Pretorius, A.; van Biljon, E.; van der Westhuizen, E.; van Staden, L.; and Kamper, H. 2019. Unsupervised Acoustic Unit Discovery for Speech Synthesis Using Discrete Latent-Variable Neural Networks. *Proc. Interspeech 2019* 1103–1107.

Godard, P.; Adda, G.; Adda-Decker, M.; Benjumea, J.; Besacier, L.; Cooper-Leavitt, J.; Kouarata, G.-N.; Lamel, L.; Maynard, H.; Müller, M.; et al. 2017. A very low resource language speech corpus for computational language documentation experiments. *arXiv preprint arXiv:1710.03501* .

Graves, A.; Fernández, S.; Gomez, F.; and Schmidhuber, J. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, 369–376. ACM.

Griffin, D.; and Lim, J. 1984. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32(2): 236–243.

Jia, Y.; Weiss, R. J.; Biadsy, F.; Macherey, W.; Johnson, M.; Chen, Z.; and Wu, Y. 2019. Direct Speech-to-Speech Translation with a Sequence-to-Sequence Model. *Proc. Interspeech 2019* 1123–1127.

Kamper, H.; Livescu, K.; and Goldwater, S. 2017. An embedded segmental k-means model for unsupervised segmentation and clustering of speech. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 719–726. IEEE.

Kuhl, P. K.; Conboy, B. T.; Coffey-Corina, S.; Padden, D.; Rivera-Gaxiola, M.; and Nelson, T. 2008. Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences* 363(1493): 979–1000.

Lample, G.; Conneau, A.; Denoyer, L.; and Ranzato, M. 2018. Unsupervised Machine Translation Using Monolingual Corpora Only. In *International Conference on Learning Representations*.

Lavie, A.; Waibel, A.; Levin, L.; Finke, M.; Gates, D.; Gavalda, M.; Zeppenfeld, T.; and Zhan, P. 1997. JANUS-III: Speech-to-speech translation in multiple languages. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, 99–102. IEEE.

Lewis, M. P.; Simons, G. F.; and Fennig, C. D. 2009. Ethnologue: languages of the world, Dallas, Texas: SIL International. *Online version: http://www. ethnologue. com* 12(12): 2010.

Li, X.; Dalmia, S.; Mortensen, D. R.; Li, J.; Black, A. W.; and Metze, F. 2020. Towards zero-shot learning for automatic phonemic transcription. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Liu, A. H.; Tu, T.; Lee, H.-y.; and Lee, L.-s. 2019. Towards Unsupervised Speech Recognition and Synthesis with Quantized Speech Representation Learning. *arXiv preprint arXiv:1910.12729* .

Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1412–1421.

Matusov, E.; Kanthak, S.; and Ney, H. 2005. On the integration of speech recognition and statistical machine translation. In *Ninth European Conference on Speech Communication and Technology*.

Muthukumar, P. K.; and Black, A. W. 2014. Automatic discovery of a phonetic inventory for unwritten languages for statistical speech synthesis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2594–2598. IEEE.

Nakamura, S.; Markov, K.; Nakaiwa, H.; Kikui, G.-i.; Kawai, H.; Jitsuhiro, T.; Zhang, J.-S.; Yamamoto, H.; Sumita, E.; and Yamamoto, S. 2006. The ATR multilingual speech-to-speech translation system. *IEEE Transactions on Audio, Speech, and Language Processing* 14(2): 365–376.

Ney, H. 1999. Speech translation: Coupling of recognition and translation. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, volume 1, 517–520. IEEE.

Oord, A. v. d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* .

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics.

Post, M.; Kumar, G.; Lopez, A.; Karakos, D.; Callison-Burch, C.; and Khudanpur, S. 2013. Improved speech-to-text translation with the Fisher and Callhome Spanish–English speech translation corpus. In *Proc. IWSLT*.

Ren, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2019. Almost Unsupervised Text to Speech and Automatic Speech Recognition. In *International Conference on Machine Learning*, 5410–5419.

Salesky, E.; Sperber, M.; and Black, A. W. 2019. Exploring phoneme-level speech representations for end-to-end speech translation. *arXiv preprint arXiv:1906.01199* .

Scharenborg, O.; Besacier, L.; Black, A.; Hasegawa-Johnson, M.; Metze, F.; Neubig, G.; Stüker, S.; Godard, P.; Müller, M.; Ondel, L.; et al. 2020. Speech technology for unwritten languages. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28: 964–975.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2019. MASS: Masked Sequence to Sequence Pre-training for Language Generation. *arXiv preprint arXiv:1905.02450* .

Sperber, M.; Neubig, G.; Niehues, J.; and Waibel, A. 2019. Attention-Passing Models for Robust and Data-Efficient End-to-End Speech Translation. *TACL* 7: 313–325.

Tjandra, A.; Sakti, S.; and Nakamura, S. 2019. Speech-to-speech Translation between Untranscribed Unknown Languages. *arXiv preprint arXiv:1910.00795* .

Tjandra, A.; Sisman, B.; Zhang, M.; Sakti, S.; Li, H.; and Nakamura, S. 2019. VQVAE Unsupervised Unit Discovery and Multi-scale Code2Spec Inverter for Zerospeech Challenge 2019 .

van den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, 6306–6315.

Vaswani, A.; Bengio, S.; Brevdo, E.; Chollet, F.; Gomez, A.; Gouws, S.; Jones, L.; Kaiser, Ł.; Kalchbrenner, N.; Parmar, N.; et al. 2018. Tensor2Tensor for Neural Machine Translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, 193–199.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*, 5998–6008.

Versteegh, M.; Anguera, X.; Jansen, A.; and Dupoux, E. 2016. The zero resource speech challenge 2015: Proposed approaches and results. *Procedia Computer Science* 81: 67–72.

Vigliocco, G.; Vinson, D. P.; Lewis, W.; and Garrett, M. F. 2004. Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive psychology* 48(4): 422–488.

Vila, L. C.; Escolano, C.; Fonollosa, J. A.; and Costa-jussà, M. R. 2018. End-to-End Speech Translation with the Transformer. In *IberSPEECH*, 60–63.

Wahlster, W. 2013. *Verbmobil: foundations of speech-to-speech translation*. Springer Science & Business Media.

Weiss, R. J.; Chorowski, J.; Jaitly, N.; Wu, Y.; and Chen, Z. 2017. Sequence-to-Sequence Models Can Directly Translate Foreign Speech. *Proc. Interspeech 2017* 2625–2629.

Wilkinson, A.; Zhao, T.; and Black, A. W. 2016. Deriving Phonetic Transcriptions and Discovering Word Segmentations for Speech-to-Speech Translation in Low-Resource Settings. In *INTERSPEECH*, 3086–3090.

Wind, J. 1989. The evolutionary history of the human speech organs. *Studies in language origins* 1: 173–197.