Pronunciation-Lexicon Free Training for Phoneme-based Crosslingual ASR via Joint Stochastic Approximation

Anonymous ACL submission

Abstract

Recently, pre-trained models with phonetic supervision have demonstrated their advantages for crosslingual speech recognition in data efficiency and information sharing across languages. However, a limitation is that a pronunciation lexicon is needed for such phoneme-based crosslingual speech recognition. In this study, we aim to eliminate the need for pronunciation lexicons and propose a latent variable model based method, with phonemes being treated as discrete latent variables. The new method consists of a speech-to-phoneme (S2P) model and a phoneme-to-grapheme (P2G) model, and a grapheme-to-phoneme (G2P) model is introduced as an auxiliary inference model. To jointly train the three models, we utilize the joint stochastic approximation (JSA) algorithm, which is a stochastic extension of the EM (expectation-maximization) algorithm and has demonstrated superior performances particularly in estimating discrete latent variable models. Based on the Whistle multilingual pre-trained S2P model, crosslingual experiments on Polish (130h) and Indonesian (20h) are conducted. With only 10 minutes of phoneme supervision, the new method, SPG-JSA, achieves 5% error rate reductions compared to the best cross-lingual fine-tuning approach using subword or full phoneme supervision. Furthermore, it is found that in language domain adaptation (i.e., utilizing cross-domain text-only data), SPG-JSA outperforms the standard practice of language model fusion via the auxiliary support of the G2P model by 9% error rate reductions.

1 Introduction

In recent years, automatic speech recognition (ASR) systems based on deep neural networks (DNNs) have made significant strides, which benefit from large amounts of transcribed speech data. Remarkably, more than 7,000 languages are spoken worldwide (Ethnologue, 2019), and most of them are low-resourced languages. A pressing challenge for the speech community is to develop ASR systems for new, unsupported languages rapidly and cost-effectively. Crosslingual ASR have been explored as a promising solution to bridge this gap (Schultz and Waibel, 1998; Conneau et al., 2021; Babu et al., 2021; Zhu et al., 2021).

In crosslingual speech recognition, a pre-trained multilingual model is fine-tuned to recognize utterances from a new, target language, which is unseen in training the multilingual model. In this way, crosslingual speech recognition could achieve knowledge transfer from the pre-trained multilingual model to the target model, thereby reducing reliance on transcribed data and becoming one of the effective solutions for low-resource speech recognition. Most recent research on pre-training for cross-lingual ASR can be classified into three categories - supervised pre-training with graphemic transcription or phonetic transcription, and self-supervised pre-training. The pros and cons of the three categories have recently been discussed in (Yusuyin et al., 2024). Under a common experimental setup with respect to pre-training data size and neural architecture, it is further found in (Yusuyin et al., 2024) that when crosslingual fine-tuning data is more limited, phoneme-based supervised pre-training achieves the most competitive results and provides high data-efficiency. This makes sense since phonetic units such as described in International Phonetic Alphabet (IPA), are exactly those sounds shared in human language throughout the world. In contrast, the methods using grapheme units face challenges in learning shared crosslingual representations due to a lack of shared graphemes among different languages.

A longstanding challenge in phoneme-based speech recognition is that phoneme labels are needed for each training utterance. Phoneme labels are usually obtained by using a manually-crafted pronunciation lexicon (PROLEXs), which maps

every word in the vocabulary into a phoneme sequence. Grapheme-to-phoneme (G2P) tools have been developed to aid this process of labeling sentences from their graphemic transcription into phonemes, but such tools are again created based on PROLEXs. There are enduring efforts to compile PROLEXs and develop G2P tools (Novak et al., 2016; Mortensen et al., 2018; Hasegawa-Johnson et al., 2020) for different languages. Overall, the existing approaches of phoneme-based ASR heavily depend on expert labor and are not scalable to be applied to much more low-resource languages.

In this paper we are interested in reducing the reliance on PROLEXs in building phoneme-based crosslingual ASR systems, i.e., towards PROLEX free. In recognizing speech x into text y, phonemes arise as intermediate states. So intuitively we propose to treat phonemes as hidden variables h, and construct a latent variable model (LVM) with pairs of speech and text (x, y) as observed values. Basically, the whole model is a conditional generative model from Speech to Phonemes and then to Graphemes, which is referred to as a SPG model, denoted by $p_\theta(h, y|x)$. SPG consists a speech-to-phoneme (S2P) model $p_\theta(h|x)$ and a phoneme-to-grapheme (P2G) model $p_\theta(y|h)$, and is thus a two-stage model. Latent variable modeling enables us to train the SPG model, without the need to knowing h, by maximizing marginal likelihood $p_\theta(y|x)$. This is different from previous two-stage ASR model with phonemes as intermediate states, as reviewed later in Section 2. Learning latent-variable models usually involves introducing an auxiliary G2P model $q_\phi(h|y)$.

Method contribution. Note that phonemes take discrete values, and recently the joint stochastic approximation (JSA) algorithm (Xu and Ou, 2016; Ou and Song, 2020) has emerged for learning discrete latent variable models with impressive performance. In this paper we propose to apply JSA to learn the SPG model, which is called the SPG-JSA approach. The S2P model is initialized from a pre-trained phoneme-based multilingual S2P backbone, called Whistle (Yusuyin et al., 2024). In practice, when viewing phonemes as labels, we combine supervised learning over 10 minutes of transcribed speech with weak phoneme labels and unsupervised learning over a much larger dataset without phoneme labels. Bootstrapping from a good S2P backbone (like Whistle) and providing few-shots samples of latent variables (such as 10 minutes of weak phoneme labels) is found to be important to make SPG-JSA successfully work in the challenging task of crosslingual ASR.

Experiment contribution. Crosslingual experiments on Polish (130h) and Indonesian (20h) are conducted. With only 10 minutes of phoneme supervision, SPG-JSA outperforms the best crosslingual fine-tuning approach using subword or full phoneme supervision. Furthermore, it is found that in language domain adaptation (i.e., utilizing cross-domain text-only data), SPG-JSA significantly outperforms the standard practice of language model fusion via the auxiliary support of the G2P model.

2 Related Work

Crosslingual ASR. Multilingual and crosslingual speech recognition has been studied for a long time (Schultz and Waibel, 1998). Modern crosslingual speech recognition typically fine-tunes a multilingual model pre-trained on multiple languages. Most recent research on multilingual pre-training can be classified into three categories - supervised pre-training with graphemic transcription (Li et al., 2021; Pratap et al., 2020; Tjandra et al., 2023; Radford et al., 2023) or phonetic transcription (Li et al., 2020; Zhu et al., 2021; Tachbelie et al., 2022; Yusuyin et al., 2023), and self-supervised pre-training (Conneau et al., 2021; Babu et al., 2021; Pratap et al., 2024). It is shown in (Yusuyin et al., 2024) that when crosslingual fine-tuning data is more limited, phoneme-based supervised pre-training can achieve better results compared to subword-based supervised pre-training and self-supervised pre-training. However, phoneme-based crosslingual fine-tuning in (Yusuyin et al., 2024) requires phoneme labels for every training utterance from the target language, which relies on a manually-crafted PROLEX for the target language. The UniSpeech method (Wang et al., 2021b) combines a phoneme-based supervised loss and a self-supervised contrastive loss to improve pre-training, and crosslingual fine-tuning still needs PROLEXs.

Two-stage ASR. The two-stage of recognizing speech to phonemes and then to graphemes has been studied for crosslingual ASR (Xue et al., 2023; Lee et al., 2023). The motivation is similar to ours that phoneme units facilitate the learning of shared phonetic representations, making cross-lingual transfer learning effective. However, both studies require a PROLEX for the target language.

Discrete latent variable models. Hidden Markov models (HMMs) are classic discrete latent

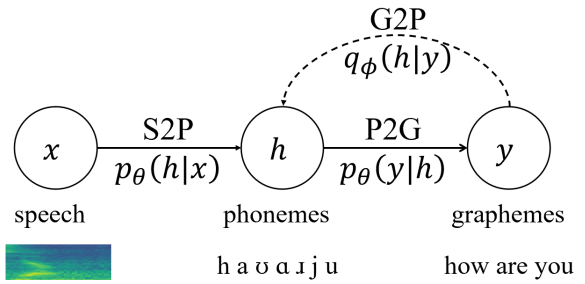

Figure 1: Overview of the latent variable model (SPG), consisting of speech-to-phoneme (S2P) and phoneme-to-grapheme (P2G). Learning SPG without knowing h involves introducing an auxiliary G2P model, denoted by the dashed line.

variable models (LVMs) and have been applied to ASR for a long time (Rabiner, 1989). Discrete LVMs are seldom used in recent end-to-end ASR systems, but has been widely used in many other machine learning applications such as dialog systems (Kim et al., 2020; Zhang et al., 2020), program synthesis (Chen et al., 2021), and discrete representation learning (van den Oord et al., 2017).

3 Method: SPG-JSA

3.1 Model

Let (x, y) denote the pair of speech and text for an utterance. Specifically, x represents the speech log-mel spectrogram and y the graphemic transcription of x. Let h denote the IPA phoneme sequence representing the pronunciation of x. In recognizing speech x into text y, we treat phonemes h as hidden variables, and construct a latent variable model, which can be decomposed as follows:

$$p_\theta(h, y|x) = p_\theta(h|x)p_\theta(y|h)$$

Basically, as shown in Figure 1, the whole model is a conditional generative model from Speech to Phonemes and then to Graphemes, which is referred to as a SPG model. SPG consists a speech-to-phoneme (S2P) model $p_\theta(h|x)$ and a phoneme-to-grapheme (P2G) model $p_\theta(y|h)$.

3.2 Training

Training the SPG model from complete data, i.e., knowing h, can be easily realized by supervised training. To train S2P and P2G end-to-end (i.e., conducting unsupervised training without knowing h), we resort to maximizing the marginal likelihood $p_\theta(y|x)$ and applying the JSA algorithm (Xu and Ou, 2016; Ou and Song, 2020), which has emerged for learning discrete latent variable models with impressive performance.

JSA involves introducing an auxiliary inference model to approximate the intractable posterior $p_\theta(h|x, y)$, which, in the ASR task considered in this paper, is assumed to take the form of $q_\phi(h|y)$, i.e., a G2P model. We can jointly train the three models (S2P, P2G and G2P), which is summarized in Algorithm 1 (SPG-JSA).

The JSA algorithm can be viewed as a stochastic extension of the well-known EM algorithm (Dempster et al., 1977), which iterates Markov Chain Monte Carlo (MCMC) sampling and parameter updating. The sampling step in JSA stochastically fills the latent variable h (phonemes) through sampling from the posterior $p_\theta(h|x, y)$, which is analogous to the E-step in EM. Particularly, using $p_\theta(h|x, y)$ as the target distribution and $q_\phi(h|y)$ as the proposal, we sample h through Metropolis independence sampler (MIS) (Liu, 2001) as follows:

1) Propose $h \sim q_\phi(h|y)$;

2) Accept h with probability $\min\left\{1, \frac{w(h)}{w(\tilde{h})}\right\}$, where

$$w(h) = \frac{p_\theta(h|x, y)}{q_\phi(h|y)} \propto \frac{p_\theta(h|x)p_\theta(y|h)}{q_\phi(h|y)} \quad (1)$$

is the usual importance ratio between the target and the proposal distribution and $\tilde{h}$ denotes the previous value for h along the Markov chain. In practice, we run MIS for several steps but for simplicity Algorithm 1 only shows a single step of MIS, with the chain is initialized from $p_\theta(h|x)$.

Once we obtain the sampled pseudo labels for h from MIS, we can treat them as if being given and calculate the gradients for the S2P, P2G, and G2P models respectively and proceed with parameter updating, similar to the process in supervised training. This is analogous to the M-step in EM, but with the proposal q_ϕ being adapted as well.

Semi-supervised Training. The SPG-JSA algorithm is general and is in fact an unsupervised learning over (x, y) without the need to know h. It is challenging to run this purely unsupervised form from scratch in the ASR task considered in this paper, which involves very high-dimensional latent space. Two additional techniques are incorporated to add inductive bias into model training. First, the S2P model is initialized from a pre-trained phoneme-based multilingual S2P backbone, called Whistle (Yusuyin et al., 2024), which have been shown to have good phoneme classification ability. In our experiments, the S2P, P2G, and G2P

Algorithm 1 The SPG-JSA algorithm

Input: S2P model $p_\theta(h|x)$, P2G model $p_\theta(y|h)$, G2P model $q_\phi(h|y)$, training dataset $\{(x, y)\}$

repeat

Draw a pair of speech and text (x, y);

Initialize $\tilde{h}$ by sampling from $p_\theta(h|x)$;

Monte Carlo sampling:

Sample h from the proposal $q_\phi(h|y)$;

Accept $\tilde{h} = h$ with probability $\min\left\{1, \frac{p_\theta(h|x)p_\theta(y|h)}{q_\phi(h|y)} / \frac{p_\theta(\tilde{h}|x)p_\theta(y|\tilde{h})}{q_\phi(\tilde{h}|y)}\right\}$;

Parameter updating:

Updating θ by ascending: $\nabla_\theta[p_\theta(\tilde{h}|x)p_\theta(y|\tilde{h})]$;

Updating ϕ by ascending: $\nabla_\phi q_\phi(\tilde{h}|y)$;

until convergence

return θ and ϕ

models are all implemented by CTC (Graves et al., 2006), which will be more detailed in Section 4.2. Second, we assume that 10 minutes of transcribed speech with phoneme labels are available, which takes much less labor than compiling a complete PROLEX for a target language. Thus, we combine supervised learning over 10 minutes speech with phoneme labels and unsupervised learning over a much larger dataset without phoneme labels. Bootstrapping from a good S2P backbone (Whistle) and providing few-shots samples of latent variables (10 minutes of phoneme labels) is found to be important to make SPG-JSA successfully work in the challenging task of crosslingual ASR.

3.3 Decoding

In testing, the S2P model first decodes out the phoneme sequence h using BeamSearch and selects the best beam as input for the P2G model. Then, the P2G model also employs BeamSearch to decode the speech recognition results, which is named as "w/o LM" result. Similar to the subword-based Whistle model (Yusuyin et al., 2024), we use an n-gram language model for WFST-based decoding, which is named as "w LM" result.

Marginal likelihood scoring. Note that the training objective of the JSA algorithm is maximizing the marginal likelihood $p_\theta(y|x)$. The decoding procedure in Section 3.3 is a crude approximation to the training objective, which is referred to as "crude decoding". So we propose a new decoding algorithm, called "decoding with marginal likelihood scoring" (MLS). It consists of the following steps: 1) S2P takes in the audio x and outputs the BeamSearch best result $\hat{h}$; 2) P2G takes in the $\hat{h}$ and generates an n-best list of candidates $\hat{y}$ using WFST decoding; 3) G2P takes in each candidate hypothesis $\hat{y}$ and propose l samples h from $q_\phi(h|\hat{y})$; 4) The marginal likelihood can be estimated with importance weights (Xu and Ou, 2016), as shown in Eq. 1; 5) Each candidate hypothesis $\hat{y}$ is rescored using a sum of the estimated marginal likelihood and the weighted LM score. In summary, the above steps can be written as:

$$y^* = \arg\max_{\hat{y}} \log \sum_{i=1}^{l} \frac{p_\theta(h_i|x)p_\theta(\hat{y}|h_i)}{q_\phi(h_i|\hat{y})} + \lambda \log P_{\text{LM}}(\hat{y}) \quad (2)$$

where $\hat{y}$ takes from the n-best list from crude decoding, and λ is LM weight. Additionally, note that crude decoding only uses the single best S2P result to fed to P2G for decoding, which is easily prone to error propagation. Decoding with MLS overcomes this drawback by scoring with multiple h.

Improving P2G via data augmentation. Note that during the SPG-JSA training, as the models gradually converge, the diversity of phoneme sequences sampled by MIS decreases. The P2G model is gradually trained with less noisy input, compared with the input fed to P2G in testing. In order to improve the robustness of the P2G model, we further augment the P2G model after the SPG-JSA training. Particularly, we decode 128 best phoneme sequences by S2P BeamSearch decoding and pair them with text labels, which serve as augmented data to further train the P2G model.

3.4 Language Domain Adaptation

Note that after SPG-JSA training, we can use the auxiliary G2P model to generate phoneme labels on pure text. Below, we take the language domain adaptation task as an example to introduce the bonus brought by the G2P model.

Text-only data is easier to obtain than transcribed speech data. In cross-domain ASR, a common approach is to train external language models for language domain adaptation. In contrast, in SPG-JSA, we can use the G2P model to generate 64 best phoneme labels through BeamSearch decoding, and then use the pairs of phonemes and text to continue adapting the P2G model. Then, we use

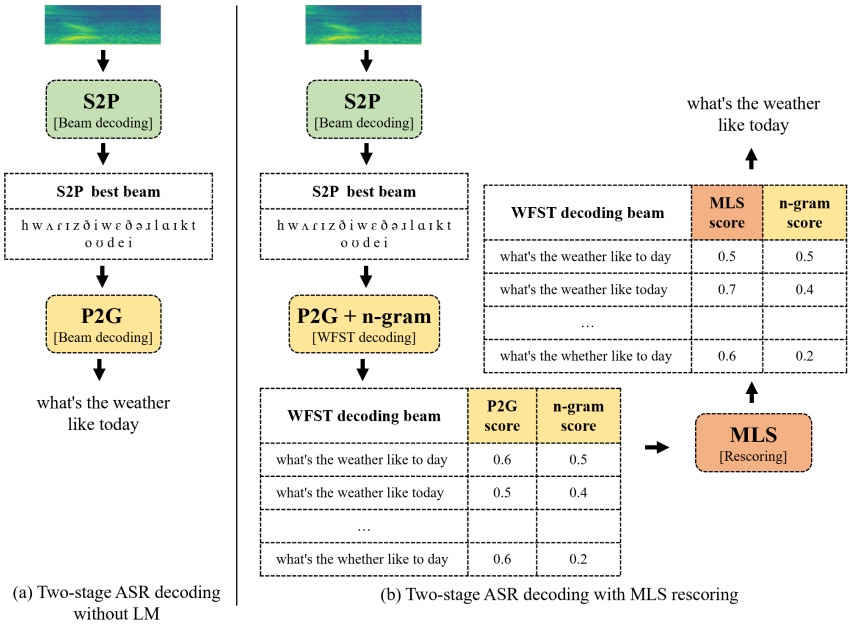

Figure 2: Illustration of decoding in SPG-JSA. (a) Decoding without LM; (b) MLS rescoring.

the original S2P, the adapted P2G, and the cross-domain language model for speech recognition on cross-domain audio, which is found to outperform the standard practice of only doing language model fusion.

4 Experiment

4.1 Datasets

Common Voice (Ardila et al., 2020) is a large multilingual speech corpus, with spoken content taken primarily from Wikipedia articles. We conduct experiments on the Common Voice dataset released at September 2022 (v11.0). We select Polish (pl) and Indonesian (id) for SPG-JSA experiments, which were not used in Whistle pre-training. Polish has 130 hours of training data, while Indonesian has 20 hours, with an average sentence length of 4.3 and 4.5 seconds, respectively. We selected 100 text sentences from the training set of each language and converted them into phonetic annotations using a publicly available phonemizer (Novak et al., 2016). In the SPG-JSA experiment, we utilized all the audio data from the two language training sets along with the corresponding text transcriptions and 100 sentences (about 10 minutes) of phonetic labels.

VoxPopuli (Wang et al., 2021a) is a multilingual speech dataset of parliamentary speeches in 23 European languages from the European Parliament. The Polish training set consists of 94.5 hours (or 710,000 words) transcribed speech data, with an average sentence length of 10 seconds. We use the training set texts for language domain adaptation experiments. Additionally, the Polish validation set is used for model selection, and the test set is used for evaluation.

Indonesian in-house data. We conducted Indonesian language domain adaptation experiments using our in-house dataset (VoxPopuli does not include Indonesian). This dataset consists of 798 hours (or 6.16 M words) transcribed speech data, with an average sentence length of 5.18 seconds. We use the training set texts for language domain adaptation experiments. Additionally, the validation set is used for model selection, and the test set is used for evaluate the experimental results.

4.2 Setup

For phoneme-based models, both of the polish and Indonesian alphabet size of phonemes is 35. For subword-based models, both of the polish and Indonesian alphabet size of subwords is 500. All text normalization and phonemization strategies are consistent with the Whistle work (Yusuyin et al., 2024). For each language, we use its transcripts to separately train a word-level n-gram language model for WFST-based decoding.

In the experiments, the S2P, P2G, and G2P models are all based on CTC. The Whistle-small 90M pre-trained model [1] is used to initialize the S2P model. Both the G2P and P2G models use 8-layer Transformer encoders with dimension 512. We set the self-attention layer to have 4 heads with

[1] https://github.com/thu-spmi/CAT/tree/master/egs/cv-lang10/exp/Multilingual/Multi._phoneme_S

512-dimension hidden states, and the feed-forward network (FFN) dimension to 1024. In the SPG-JSA training, for every data item, we obtain 10 samples from G2P and run MIS for each sample. All experiment are taken with the CAT toolkit (An et al., 2020). The learning rate for SPG-JSA is set to 3e-5, and when the validation loss does not decrease 10 epochs, the learning rate is multiplied by 0.5, training stop until it reaches 1e-6. We extract 80-dimension FBank features from audio (resampled to 16KHz) as inputs to the S2P model. A beam size of 16 is used for S2P and P2G decoding in testing. We average the three best-performing checkpoints on the validation set for testing.

5 Result and Analysis

5.1 SPG-JSA Results

Baseline results are taken from (Yusuyin et al., 2024), including monolingual phoneme-based training and subword-based training. The phoneme-based training utilized full phonetic annotations for 130 hours of Polish and 20 hours of Indonesian data. The phoneme-based Whistle-small pretrained model were further fine-tuned with either phoneme labels or subword labels for crosslingual speech recognition, which correspond to "Whistle phoneme FT" and "Whistle subword FT" in Table 1 and represent the two state-of-the-art cross-lingual fine-tuning approaches. Phoneme fine-tuning used full phonetic annotations.

Remarkably, in Indonesian, "Whistle phoneme FT" outperforms "Whistle subword FT", whereas in Polish, the opposite is observed. As analyzed in (Yusuyin et al., 2024), when crosslingual fine-tuning data are more limited (Indonesian has 20 hours of data vs Polish 130 hours), phoneme-based fine-tuning is more data-efficient and performs better than subword fine-tuning.

In the following, we introduce the SPG crosslingual ASR experiments with only 10 minutes of data per language having phoneme annotations. The SPG experiments have two settings – "SPG init from G2P" and SPG-JSA, as shown in Table 1, characterized by different initialization schemes. In SPG-JSA, we first fine-tune the Whistle model on 10 minutes of phoneme labels to initialize the S2P model. Subsequently, this S2P model is utilized to generate phoneme pseudo-labels on the full training set, which are then used to train the P2G and G2P models for initialization. Ater such a initialization, the SPG-JSA algorithm is employed to jointly train the three models (S2P, P2G and G2P).

Alternatively, the G2P model (instead of the S2P model) can first be fine-tuned on 10 minutes of phoneme labels and used to generate phoneme pseudo labels on the full training set, which are then be used to train S2P and P2G models. The resulting S2P, P2G and G2P model are referred to as "SPG init from G2P". A further application of the SPG-JSA algorithm to jointly train the three models is found to bring no improvement and so we only report the result for "SPG init from G2P". Presumably, this is because the G2P fine-tuned in this way is overfit to the 10 minutes of phoneme labels, which is difficult to be further trained to be a good proposal.

For the results in Table 1, SPG-JSA consistently achieves 5% error rate reductions compared to the best cross-lingual fine-tuning approach using subword or full phoneme supervision (3.82 vs 3.64 for Polish, 2.43 vs 2.31 for Indonesian). Compared to "Whistle phoneme FT", the training of the three models in SPG-JSA is trained in a more end-to-end way by maximizing marginal likelihoods over pairs of speech and text, hence obtaining better performance. Full supervision of weak phoneme labels produces worse results, presumably because the full phoneme annotations may contain errors. "Whistle subword FT" is trained in a end-to-end way. The inferior results of "Whistle subword FT" compared to SPG-JSA indicates the advantage of the SPG architecture (the benefit of explicitly modeling phonemes). The fine-tuning of Whistle using subword labels breaks the structure imposed by SPG. Using phonemes as an interface between speech and language in the SPG architecture enforces a useful structure. A seemingly problem is that the SPG pipeline contains descretized phonemes and gradients cannot propagate from P2G to S2P. However, the optimization of SPG-JSA is actually in an end-to-end way – JSA is powerful in learning discrete latent variable models, propagating gradients across the P2G and S2P components, which is different from optimizing a single neural network but turns out to be very effective.

"SPG init from G2P" shows worse results than SPG-JSA. Note that in decoding, it is the hypothesized phonemes from S2P that are fed to P2G. But in the training procedure in "SPG init from G2P", the pseudo phoneme labels used to train P2G are from G2P. This causes some mismatch in training and decoding, and presumably explains the worse

Exp.	Polish				Indonesian			
	PER	w/o LM	w LM	*MLS*	PER	w/o LM	w LM	*MLS*
Monolingual phoneme †	2.82	NA	4.97	NA	5.74	NA	3.28	NA
Monolingual subword †	NA	19.38	7.12	NA	NA	31.96	10.85	NA
Whistle phoneme FT †	1.97	NA	4.30	NA	4.79	NA	2.43	NA
Whistle subword FT †	NA	5.84	3.82	NA	NA	12.48	2.92	NA
SPG init from G2P	17.72	8.73	4.68	5.91	21.85	10.15	3.81	3.09
+ P2G augmentation	17.72	5.93	4.97	5.88	21.85	6.34	3.44	2.91
SPG-JSA	17.35	8.19	4.65	3.93	20.66	9.04	3.26	2.47
+ P2G augmentation	17.35	4.64	4.37	**3.64**	20.66	4.55	2.92	**2.31**

Table 1: PERs (%) and WERs (%) for SPG-JSA experiment on Common Voice dataset. FT: fine-tuning. MLS: marginal likelihood scoring. † denotes results from (Yusuyin et al., 2024). NA denotes not applied.

result of "SPG init from G2P".

5.2 Language Domain Adaptation Results

As shown in Table 2, for Polish, we test our models on VoxPopuli Polish test set, while both Whistle and SPG-JSA models is train on the Common Voice dataset. The CommonVoice dataset is comprised of texts from Wikipedia, recorded by users on mobile devices, while the VoxPopuli dataset consists of audio recordings of speeches from the European Parliament. Notably, 61.5% of the words in the VoxPopuli Polish training set do not appear in the CommonVoice vocabulary list, and 31.5% of the words in the test set are also absent. This indicates significant differences between the two datasets in terms of linguistic context, vocabulary, recording equipment, and average sentence length. We only use the text from the VoxPopuli training set and train a word-level 4-gram language model for language model fusion.

The first row at Table 2 shows the results of testing Whistle subword fine-tuning model directly with cross-domain language model integration, which is a common method used in cross-domain speech recognition. We then test SPG-JSA model directly without further training (the second row of Table 2). Comparing the two result on Polish reveals that the SPG-JSA model performs better on cross-domain ASR tasks, indicating its stronger robustness. We further apply the domain adaptation method, introduced in Section 3.4, to continue training the P2G model on VoxPopuli training text. The result clearly demonstrates the advantage of SPG-JSA, and its performance far exceeds that of traditional language domain adaptation method by 9% error rate reduction (22.58 vs 20.57).

For Indonesian, our in-house Indonesian dataset is from audio books, which has a clear domain difference from the CommonVoice dataset. Indonesian experiments are taken simlarly to Polish. The SPG-JSA model with the adapted P2G model obtains the best result, significantly outperform Whistle subword fine-tuning model by 9% error rate reduction (12.39 vs 11.23) as well.

5.3 Analysis and Ablation

To provide an intuitive understanding of the SPG-JSA training process, Figure 3 in Appendix A shows the changes in several key indicators over the number of training iterations. It can be seen that the training losses of all three models and the validation error rates gradually decrease when using the SPG-JSA Algorithm 1, clearly showing the ability of SPG-JSA for model optimization. Through SPG-JSA training, compared to the model fine-tuned with only 10 minutes of phonetic labels, which is the initial model in the experiment, the SPG-JSA model achieves a relative PER reduction of 45% and a WER reduction of 48% on the validation set.

Table 3 shows ablation experiments with different amounts of supervised phoneme data. As the amount of supervised data increases, both PER and WER of the SPG-JSA model significantly decrease. Compared to 2 minutes of supervised data, with 10 minutes semi-supervised training, PER decreases by 36% and WER by 8% in polish; Compared to unsupervised training, PER decreases by 65% and WER by 18%. There is also the same trend in the semi-supervised SPG-JSA experiments of the Indonesian.

Exp.	Polish			Indonesian		
	w/o LM	w LM	*MLS*	w/o LM	w LM	*MLS*
Whistle subword FT on CV	33.46	22.58	NA	43.69	12.39	NA
SPG-JSA on CV	35.18	29.04	26.79	39.19	16.93	14.28
+ LDA training	28.87	23.84	**20.57**	30.69	12.68	**11.23**

Table 2: WERs (%) of cross-domain language domain adaptation (LDA) experiments from CV to VoxPopuli Polish and in-house Indonesian datasets. The FT denotes fine-tuning. The MLS denotes marginal likelihood scoring.

Amount of supervised data	Polish				Indonesian			
	PER	w/o LM	w LM	*MLS*	PER	w/o LM	w LM	*MLS*
unsupervised	50.24	13.43	6.20	5.05	32.71	11.09	3.74	2.80
+ P2G augmentation	50.24	6.08	5.28	4.48	32.71	5.33	3.28	2.47
20 sentences (about 2 minutes)	27.35	8.25	5.17	4.25	27.37	9.01	3.39	2.66
+ P2G augmentation	27.35	5.31	4.78	3.96	27.37	5.45	3.04	2.47
100 sentences (about 10 minutes)	**17.35**	8.19	4.65	3.93	**20.66**	9.04	3.26	2.47
+ P2G augmentation	**17.35**	**4.64**	**4.37**	**3.64**	**20.66**	**4.55**	**2.92**	**2.31**

Table 3: Performance comparison of different amounts of phoneme labels as supervised data in SPG-JSA training. MLS: marginal likelihood scoring.

On the other hand, with reduced amounts of phoneme supervision such as only 2 minutes or even zero-shot, SPG-JSA obtains impressive results. Notably, the zero-shot PERs of Polish and Indonesian by the Whistle-small model are 58% and 46% respectively. Without any phoneme supervision data, unsupervised SPG-JSA training leads to a significant reduction in PERs for both languages: a 13% decrease for Polish and a 30% decrease for Indonesian. The MLS decoding result of Indonesian (2.47%) even surpasses that of subword fine-tuning (2.92%) and approaches the result of phoneme fine-tuning (2.43%). It should be emphasized that all 35 phonemes of the Indonesian are present in the Whistle phoneme set. We copy the corresponding weights for parameter initialization in the SPG-JSA training. For Polish, 31 phonemes appear in the Whistle phoneme set, while four phonemes do not. We randomly initialized the weights for the unseen phonemes. This may account for the less-than-ideal performance of unsupervised Polish training. However, with the addition of 2 minutes of phoneme labels as supervision data, the WER from Polish SPG-JSA training (3.96%) is lower than that of 130-hour full phoneme fine-tuning (4.30%) and close to the result of subword fine-tuning (3.82%).

6 Conclusions

In this paper, we aim to achieve crosslingual speech recognition based on phonemes without pronunciation lexicons. By treating phonemes as discrete latent variables, modeling S2P and P2G together as a latent variable model (SPG), and introducing a G2P model as an auxiliary inference model, we utilize the JSA algorithm to jointly train these three networks. We refer to this new approach as SPG-JSA. Particularly, the S2P model is initialized from Whistle, a pre-trained phoneme-based multilingual S2P backbone. This paper also proposes marginal likelihood scoring and P2G augmentation, which further improve the performance of SPG-JSA. In crosslingual experiments with two languages (Polish and Indonesian), SPG-JSA method demonstrates remarkable performance. By utilizing merely 10 minutes of speech with phoneme labels, it outperforms full data (130 and 20 hours) phonetic supervision. This effectively eliminates the necessity of using a PROLEX. Moreover, the SPG-JSA method surpasses crosslingual subword fine-tuning, and we take the language domain adaptation task as an example to show the bonus brought by SPG-JSA. It is found that SPG-JSA significantly outperforms the standard practice of language model fusion via the auxiliary support of the G2P model.

Limitations

This paper presents some promising results along the SPG-JSA approach. There are some limitations with this work. First, the SPG-JSA method can be applied in phoneme-based pre-training to exploit a larger amount of data from more languages, even those languages without PROLEXs. Second, SPG-JSA simultaneously produces S2P, P2G and G2P models, each of which potentially can be applied to a variety of speech processing tasks. Language domain adaptation is only one example. G2P is useful for text-to-speech. S2P may provide a new way for discrete tokenization for speech. Third, the Whistle based S2P is multilingual; this work is limited in not building multilingual P2G and G2P.

References

Keyu An, Hongyu Xiang, and Zhijian Ou. 2020. CAT: A CTC-CRF based ASR toolkit bridging the hybrid and the end-to-end approaches towards data efficiency and low latency. In *Interspeech*, pages 566–570.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Twelfth Language Resources and Evaluation Conference*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, and 1 others. 2021. XLS-R: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech*, pages 2278–2282.

Xinyun Chen, Dawn Song, and Yuandong Tian. 2021. Latent execution for neural program synthesis beyond domain-specific languages. *Advances in Neural Information Processing Systems*.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech*, pages 2426–2430.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39.

Ethnologue. 2019. Languages of the world. https://www.ethnologue.com/.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *International Conference on Machine Learning*.

Mark Hasegawa-Johnson, Leanne Rolston, Camille Goudeseune, Gina-Anne Levow, and Katrin Kirchhoff. 2020. Grapheme-to-phoneme transduction for cross-language ASR. In *International Conference on Statistical Language and Speech Processing*.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. In *International Conference on Learning Representations (ICLR)*.

Wonjun Lee, Gary Geunbae Lee, and Yunsu Kim. 2023. Optimizing two-pass cross-lingual transfer learning: Phoneme recognition and phoneme to grapheme translation. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Bo Li, Ruoming Pang, Tara N Sainath, Anmol Gulati, Yu Zhang, James Qin, Parisa Haghani, W Ronny Huang, Min Ma, and Junwen Bai. 2021. Scaling end-to-end models for large-scale multilingual ASR. In *IEEE Automatic Speech Recognition and Understanding Workshop*, pages 1011–1018.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, and 1 others. 2020. Universal phone recognition with a multilingual allophone system. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8249–8253.

Jun S Liu. 2001. *Monte Carlo strategies in scientific computing*, volume 10. Springer.

David R Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Eleventh International Conference on Language Resources and Evaluation*.

Josef Robert Novak, Nobuaki Minematsu, and Keikichi Hirose. 2016. Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework. *Natural Language Engineering*.

Zhijian Ou and Yunfu Song. 2020. Joint stochastic approximation and its application to learning discrete latent variable models. In *Conference on Uncertainty in Artificial Intelligence*, pages 929–938. PMLR.

Vineel Pratap, Anuroop Sriram, Paden Tomasello, Awni Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert. 2020. Massively multilingual ASR: 50 languages, 1 model, 1 billion parameters. In *Interspeech*, pages 4751–4755.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi,

and 1 others. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25:1–52.

Lawrence R Rabiner. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*.

T. Schultz and A. Waibel. 1998. Multilingual and crosslingual speech recognition. In *Proc. DARPA Workshop on Broadcast News Transcription and Understanding*, pages 259–262.

Martha Yifiru Tachbelie, Solomon Teferra Abate, and Tanja Schultz. 2022. Multilingual speech recognition for GlobalPhone languages. *Speech Communication*.

Andros Tjandra, Nayan Singhal, David Zhang, Ozlem Kalinli, Abdelrahman Mohamed, Duc Le, and Michael L. Seltzer. 2023. Massively multilingual ASR on 70 languages: tokenization, architecture, and generalization capabilities. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *NIPS*.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021a. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003.

Chengyi Wang, Yu Wu, Yao Qian, Kenichi Kumatani, Shujie Liu, Furu Wei, Michael Zeng, and Xuedong Huang. 2021b. Unispeech: Unified speech representation learning with labeled and unlabeled data. In *International Conference on Machine Learning*, pages 10937–10947.

Haotian Xu and Zhijian Ou. 2016. Joint stochastic approximation learning of Helmholtz machines. In *ICLR Workshop Track*.

Hongfei Xue, Qijie Shao, Peikun Chen, Pengcheng Guo, Lei Xie, and Jie Liu. 2023. TranUSR: Phoneme-to-word transcoder based unified speech representation learning for cross-lingual speech recognition. In *INTERSPEECH*.

Saierdaer Yusuyin, Hao Huang, Junhua Liu, and Cong Liu. 2023. Investigation into phone-based subword units for multilingual end-to-end speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5.

Saierdaer Yusuyin, Te Ma, Hao Huang, Wenbo Zhao, and Zhijian Ou. 2024. Whistle: Data-efficient multilingual and crosslingual speech recognition via weakly phonetic supervision. In *arXiv*, page 2406.02166.

Yichi Zhang, Zhijian Ou, Min Hu, and Junlan Feng. 2020. A probabilistic end-to-end task-oriented dialog model with latent belief states towards semi-supervised learning. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chengrui Zhu, Keyu An, Huahuan Zheng, and Zhijian Ou. 2021. Multilingual and crosslingual speech recognition using phonological-vector based phone embeddings. In *IEEE Automatic Speech Recognition and Understanding Workshop*, pages 2301–2312.

A The training curves of SPG-JSA

The training curves of SPG-JSA are shown in Fig 3.

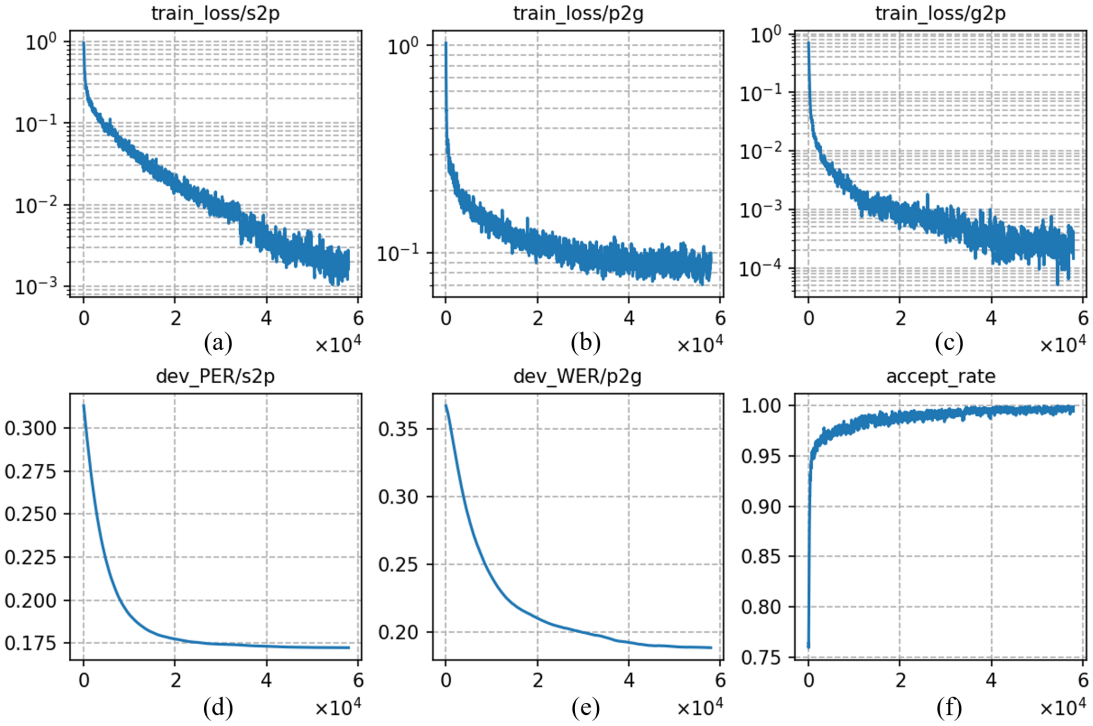

Figure 3: Plots of training and validation curves in SPG-JSA training on Common Voice polish data. (a), (b), (c) represent the train losses of the S2P, P2G, and G2P models in the SPG-JSA training, respectively. (d) and (e) are the error rates of S2P and P2G models in the validation set. (f) represents the ratio of the number of samples accepted by the MIS sampler to the total number of samples proposed by G2P in one iteration.