# Massively Multilingual Adversarial Speech Recognition

**Oliver Adams, Matthew Wiesner, Shinji Watanabe, and David Yarowsky**
Department of Computer Science
Johns Hopkins University, Baltimore MD, USA
`{oadams1,wiesner,shinjiw,yarowsky}@jhu.edu`

## Abstract

We report on adaptation of multilingual end-to-end speech recognition models trained on as many as 100 languages. Our findings shed light on the relative importance of similarity between the target and pretraining languages along the dimensions of phonetics, phonology, language family, geographical location, and orthography. In this context, experiments demonstrate the effectiveness of two additional pretraining objectives in encouraging language-independent encoder representations: a context-independent phoneme objective paired with a language-adversarial classification objective.

## 1 Introduction

The main difficulty in creating automatic speech recognition (ASR) systems for a large number of the world's 7,000 languages is a lack of training data. Such data comes in the form of speech paired with transcriptions, a pronunciation lexicon, and text for language model training. A common technique in data-constrained settings is to learn language-independent representations of speech via multilingual training. Popular approaches include the use of multilingual bottleneck features (Vesely et al., 2012) as well as multilingual model training before fine-tuning to a *target* language (Scanzio et al., 2008; Vu et al., 2012).

Prior work in multilingual and cross-lingual speech recognition has been restricted to a small handful of the world's most-spoken languages, relying on multilingual corpora such as Global-Phone (Schultz, 2002), the IARPA Babel corpora (Gales et al., 2014), or the VoxForge[1] corpora. Most work typically only reports on models trained on a subset of these languages.

In this paper we explore pretraining multilingual ASR models using speech from as many as 100 languages from the CMU Wilderness Multilingual Speech Dataset (Black, 2019).[2] To the best of our knowledge, this is the greatest number of languages that has been used in multilingual ASR model training to date. We perform experiments to guide the choice of languages used when pretraining the model and assess the relative importance of similarity between the pretraining languages and target language in terms of geographic location, phonology, phonetic inventory, language family and orthography.

We examine these variables in the context of two experimental setups: one where models are adapted to target language and target speakers, and one where models are adapted to target language but non-target speakers. The first task is relevant to language documentation contexts, which often involves transcribing speech of specific speakers for which there already exists some transcribed speech as training data (Michaud et al., 2018). The second case is relevant to incident response as modelled by LORELEI (Strassel and Tracey, 2016), where there may only be a single target-language consultant available for which transcribed speech can be elicited, but the goal is to have an ASR model that generalizes to multiple speakers.

Multilingual ASR training on such a scale presents challenges because of this language diversity. In order to guide the model to learn language-independent representations that are more amenable to adaptation, we experiment with two auxiliary training tasks. The first is context-independent phoneme sequence prediction to help bridge orthographic inconsistencies between languages. The second is a domain-adversarial classification objective (Ganin et al., 2016) over languages to encourage invariance

---

[1] `voxforge.org`

[2] `festvox.org/cmu_wilderness/index.html`

of the model with respect to language-specific phenomena. The hierarchical combination of grapheme and phoneme objectives has only been used in monolingual end-to-end frameworks (Krishna et al., 2018; Rao and Sak, 2017). Language-adversarial training in ASR (Yi et al., 2018) has not been done at this scale before, nor in an end-to-end framework.

Our experiments are designed to answer the following questions:

1. Is there benefit in scaling multilingual model training to a large number of languages?

2. In what circumstances, if any, does the addition of a phoneme and/or language-adversarial objective improve multilingual models?

3. How should we choose languages with which to pretrain a multilingual model?

4. Do the answers to the above questions change when adapting to target versus non-target speakers in the target language?

We find that using the auxiliary objectives in pretraining facilitates model transfer to unseen languages, especially when the pretraining languages are very dissimilar (Section 6). When the target speakers are seen in adaptation (Section 7), similarity of the pretraining languages and the target language is more important than quantity of pretraining languages. Choosing as pretraining languages geographically proximal languages tends to help more than phonetically and phonologically similar but otherwise distant languages. However, when adapting to a handful of non-target speakers of the target language (Section 8), the domain mismatch caused by the unseen speaker, language, or recording environment degrades performance. Exposing the model to as many pretraining languages as possible becomes vital to minimize this mismatch. Results on this task demonstrate that a massively multilingual seed model substantially outperforms other seed models trained on languages similar to the target. We have provided an ESPnet recipe to train and test our models.[3]

## 2 Related Work

This paper builds on work on multilingual ASR, end-to-end ASR, and adversarial learning.

**Multilingual transfer** in ASR often relies on using bottle-neck features (Vesely et al., 2012; Vu et al., 2012; Karafiát et al., 2018) and adapting an acoustic model trained on one language to effectively recognize the sounds of other languages (Schultz and Waibel, 2001; Le and Besacier, 2005; Stolcke et al., 2006; Tóth et al., 2008; Plahl et al., 2011; Thomas et al., 2012; Imseng et al., 2014; Do et al., 2014; Heigold et al., 2013; Scharenborg et al., 2017). However, while most work uses less than 10 languages for model training, we include up to 100 languages in training.

**End-to-end ASR** has recently become popular, with approaches such as attention-based encoder-decoder models (Chorowski et al., 2015; Chan et al., 2015), the connectionist temporal classification (CTC) objective of Graves et al. (2006, 2013), or a combination of both (Kim et al., 2016; Hori et al., 2017). These approaches have also been deployed in multilingual settings (Toshniwal et al., 2017; Chiu et al., 2018; Müller et al., 2017; Dalmia et al., 2018; Watanabe et al., 2017a). Our baseline approach to multilingual knowledge transfer is most similar to Inaguma et al. (2018), and involves training a hybrid CTC-attention seed model.

**Hierarchical and multi-task** approaches including combining grapheme and phoneme prediction in monolingual contexts (Rao and Sak, 2017; Krishna et al., 2018) at different levels of the network, or using sub-word units of varying granularity (Sanabria and Metze, 2018), have been shown to improve ASR performance. In this paper we extend the approach of hierarchical placement of additional objectives in order to enforce language independent, transferable models.

**Domain-adversarial training** is one such method for encouraging the model to learn language independent representations. A key contribution of this paper is the use of a domain-adversarial classification objective (Ganin et al., 2016) over many languages in order to encourage the model to learn representations that are invariant to language. Domain-adversarial training incorporates an auxiliary domain classification task, but negates gradients for encoder weights before the parameter update in order to guide the encoder to produce hidden representations that fool the classifier: i.e. they minimize information about the language while still facilitating the

primary task of speech recognition.

Domain-adversarial training has been used in speech recognition to learn features invariant to noise conditions (Shinohara, 2016), accents (Sun, 2018), and sex (Tripathi et al., 2018). Most closely related to our work is that of Yi et al. (2018), who use a language-adversarial objective when preparing multilingual bottleneck features from four languages for a hidden Markov model (HMM) ASR pipeline. In contrast, our work uses an adversarial objective across many languages, pairing it with a context-independent phoneme objective in an end-to-end framework.

## 3 Data

We scraped the data that forms the CMU Wilderness dataset, using a freely available script.[4] This dataset consists of dramatized readings of the Bible in hundreds of languages. Each reading is ascribed a rating based on alignment quality which fits into one of these classes: `very good`, `good`, `okay`, and `not okay`.

The script used to preprocess the data uses a universal pronunciation module in Festival (Taylor et al., 1998)[5] to produce pronunciation lexicons using an approach based on that of UniTran (Yoon et al., 2007), which we use to create phonemic transcriptions.

### 3.1 Characteristics of the Speech

The dataset consists of readings of the Bible, with readings typically of just a few speakers, mostly male. These are often dramatized, with sound effects and background music. For many purposes this could be considered a limitation of the data. Although the characteristics of the speech are unique, it allows us to investigate multilingual models over many languages without the confounds of an overly noisy environment. It is not unreasonable to expect our findings to generalize to other speech recognition domains.

### 3.2 Evaluation Languages

While the dataset includes only a single reading of the Bible for most languages, there are a number with two or more. We evaluate on languages for which we can find two or more readings. This is so that we can compare adaptation to a target

---

|  | Hours:minutes/quality per reading | | |
|---|---|---|---|
| Aymara (ayr) | 16:19/G | 18:37/G | - |
| SB Quechua (quh) | 27:41/G | 20:02/G | - |
| Kekchi (kek) | 19:32/G | 18:30/G | - |
| Ixil (ixl) | 35:06/VG | 25:35/G | 18:29/G |
| Malagasy (mlg) | 12:29/NO | 15:52/O | 15:59/G |
| Indonesian (ind) | 19:01/G | 21:20/G | 30:34/G |
| Garap (kia) | 15:34/G | 12:17/VG | - |
| Swedish (swe) | 15:55/G | 16:46/VG | - |
| Spanish (spn) | 16:35/G | 15:19/G | - |

Table 1: The duration of each reading in the evaluation languages (ISO 639-3 language codes in parentheses), before our preprocessing. Alignment quality categories are `very good` (VG), `good` (G), `okay` (O), `not okay` (NO). *SB Quechua* denotes South Bolivian Quechua.

language but not the speakers of the target reading (we refer to this task as *language adaptation*, as explored in Section 8) with adaptation to the target language as well as the target reading (we refer to this task as *reading adaptation*). We additionally restricted the evaluation languages to those that have at least one `good` or `very good` reading in terms of alignment quality. Table 1 presents the evaluation languages and readings grouped by family or geographic location, along with their durations.

## 4 Auxiliary Training Objectives

In addition to scaling ASR training to 100 languages, a key contribution of our work is the use of a context-independent phoneme objective paired with a language-adversarial classification objective in a end-to-end grapheme-based neural network, as illustrated in Figure 1.

### 4.1 Baseline Model

Our experiments are conducted within the framework of a hybrid CTC-attention end-to-end neural model using ESPnet (Watanabe et al., 2017b), which uses an encoder-decoder architecture implemented in PyTorch (Paszke et al., 2017). The encoder we use consists of VGG-like convolution layers (Simonyan and Zisserman, 2014; Sercu et al., 2016) followed by a multilayer bidirectional long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997). The decoder uses location-based attention (Chorowski et al., 2015) and an LSTM. In addition to the attention, the decoder also incorporates CTC

probabilities over graphemes to encourage monotonicity in decoding.

## 4.2 Phoneme Objective

The end-to-end neural model performs direct grapheme prediction without recourse to a pronunciation lexicon as traditional hybrid HMM-DNN models do. Since different orthographies may be mutually disjoint or only weakly related to the phonetic content of the input speech, we use a context-independent phoneme CTC objective to encourage learning of representations independent of such orthographic idiosyncrasies.

We performed limited preliminary experiments to determine how best to use the phoneme objective, which corroborated recent work in hierarchical training objectives that supports inserting the phoneme objective in the layers below the final layer (Krishna et al., 2018). We also found that using the phoneme objective during adaptation was harmful and therefore in all reported experiments we use it only during multilingual pretraining.

## 4.3 Language-Adversarial Pretraining

For language-adversarial training we used a log-linear classifier over all languages seen in pretraining. An utterance-level mean of the penultimate encoder layer states is fed into the classifier. For each batch in training we update the network using the interpolated grapheme and phoneme objectives before a separate update step using the adversarial objective.

We follow the learning rate scheduling of Ganin et al. (2016), where the weight of the adversarial objective relative to the speech recognition tasks follows $\lambda(p) = \frac{2}{1+\exp(-10p)} - 1$ over the course of training, where $p \in [0, 1]$ is a measure of training progress. We drop the adversarial objective during target language adaptation.

## 5 Experimental Setup

### 5.1 Language Versus Reading Adaptation

We chose as target adaptation languages those languages for which we have multiple readings of the Bible. This allows us to assess adaptation of the pretrained multilingual model in two scenarios: *language adaptation* and *reading adaptation*. In *reading adaptation*, it is adapted to data from each reading of the target language, including the reading from which we select held-out evaluation utterances. In *language adaptation* it is adapted only
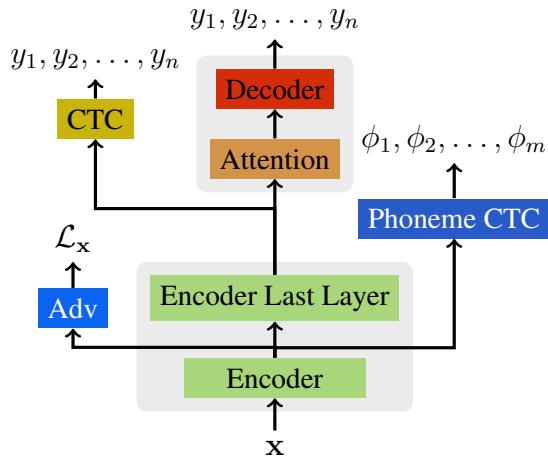


Figure 1: The end-to-end architecture used during pretraining. $\mathbf{x}$ is the input speech features, $y_1, y_2, \ldots, y_n$ is a character sequence the model is trained to output (eg. "knife"). $\phi_1, \phi_2, \ldots, \phi_m$ is a phoneme sequence the model is trained to output (eg. /naɪf/), and $\mathcal{L}_{\mathbf{x}}$ is the language identity of the input speech $\mathbf{x}$.

to readings that are not represented in the evaluation set. This last case, of adapting to just one or several speakers of a new language (in order to ultimately have a system that generalizes beyond those speakers in the language) is not common in speech recognition experimentation. Results and findings for language adaptation will be presented in Section 8.

### 5.2 Training Settings

We established training, validation and test sets for each reading using a random 80/10/10 split. When pretraining or adapting the multilingual systems, we used the combined training sets of the constituent readings.

We used 80-dimensional log Mel filterbank features with 3-dimensional pitch features. We tuned hyperparameters for these models using one Aymara reading.[6] We found that a 4 layer encoder, 1 layer decoder with 768 for the encoder hidden size and projections, decoder hidden size, and attention hidden size yielded equal-best results with deeper models. These settings were then used for training the models used in our experiments.

For the training objective, we linearly interpolated the the attentional decoder cross-entropy loss with the grapheme CTC and phoneme CTC objectives. Equal weight was given to all three since we found that to be effective in preliminary experiments. Note however, that the effective weight of

---

[6]CMU Wilderness reading ID: AYMSBU.

| Target | MONO | QUE | | CYR | | QUE+CYR | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | - | +phn+adv | - | +phn+adv | - | +phn | +adv | +phn+adv |
| Aymara | 40.6 | 34.3 | 34.5 (+0.6%) | 37.9 | 35.9 (-5.3%) | 34.6 | **34.2** | 34.8 | **34.2** (-1.2%) |
| SB Quechua | 14.8 | **13.8** | 14.0 (+1.4%) | 16.3 | 17.0 (+4.3%) | 14.9 | 14.2 | 14.0 | 13.9 (-6.7%) |
| Indonesian | 14.9 | 15.1 | 15.3 (+1.3%) | 16.1 | 17.9 (+11.2%) | 15.8 | 15.6 | 15.5 | **14.7** (-7.0%) |
| | | | Avg. rel. Δ: (+1.1%) | | Avg. rel. Δ: (+3.4%) | | | | Avg. rel. Δ: (-4.9%) |

Table 2: Word error rate (%) comparison of multilingual models adapted to target languages, with and without auxiliary training objectives (relative change in parentheses). Additionally including Cyrillic-script languages in pretraining (CYR) doesn't consistently improve over a model pretrained on Quechuan languages (QUE) unless additional phoneme and language-adversarial objectives (+phn and +adv) are used in combination (+phn+adv). The auxiliary objectives help when pretraining languages are varied, but hinder when they are very similar. The final four columns suggest that the objectives are complementary. Average relative word error rate change for each pretraining set when adding in the auxiliary objectives (versus no aditional objectives) is indicated by *Avg. rel.* Δ.

the adversarial objective effectively changes over the course of training because of the learning rate scheduling mentioned in §4.3. We trained for 15 epochs in all cases except where otherwise noted.

Note that during adaptation we initialize the model using both the multilingual encoder and decoder. We found this to work best in preliminary experimentation on a Spanish reading.

# 6 Preliminary Investigation of the Auxiliary Objectives

In this section we evaluate the use of the auxiliary phoneme and language-adversarial objectives described in Section 4 on two divergent groups of languages that are distinct along a number of dimensions, including orthography, language family and phonology, in order to assess the auxiliary objectives' capacity to bridge the divide between these languages during pretraining. This serves as an initial exploration before further experiments in Section 7 and Section 8, where we choose from a broader set of pretraining languages.

**Pretraining languages** We pretrained models on two groups of languages separately and together. The first consists of six languages from the Quechuan language family, including sub-varieties of Quechua I and II (qub, quf, qvs, qvw, qwh and qvh). We henceforth refer to this group as QUE. The second consists of six languages that use the Cyrillic script and we refer to this group as CYR. These languages include Nogai (nog), Bashkir (bak), Gagauz (gag), Khakas (kjh), Crimean Tatar (crh), and Russian (rus). With the exception of Russian, these languages are all Turkic. The character sets do not overlap between QUE and CYR and this was a deliberate choice in

this preliminary experiment to maximize the differences between the two groups.

**Evaluation languages** To test the pretrained models in varied contexts, we evaluate our models on three languages: Central Aymara (ayr), South Bolivian Quechua (SB Quechua; quh), and Indonesian (ind). These languages vary in a number of dimensions: SB Quechua is very closely related to QUE, while Indonesian is distant; Aymara is phonologically very similar to Quechuan languages, but is considered to be from a different family; Aymara had a high monolingual baseline error rate, while the others are lower; and Indonesian has three readings while the others have two. However, all evaluation languages use the Latin script. Note that in this section we assess performance in the reading adaptation case, while Section 8 presents results on the held-out reading case.

**Experiments** Table 2 compares the performance of monolingual target-language models to models adapted to the target language after being pretrained on QUE, CYR and their combination, QUE+CYR. CYR pretraining underperforms pretraining with QUE for all evaluation languages likely due to the orthographic mismatch with all of the evaluation languages. The model pretrained on QUE+CYR also underperforms QUE. Introducing the auxiliary phoneme and language-adversarial objectives helps to overcome this performance loss, making the QUE+CYR-pretrained model the best for adaptation to Aymara and Indonesian. QUE remained the best pretraining set for adaptation to SB Quechua, which is unsurprising given how well represented SB Quechua is by the languages included in the Quechuan language group. This suggests that when a substantial
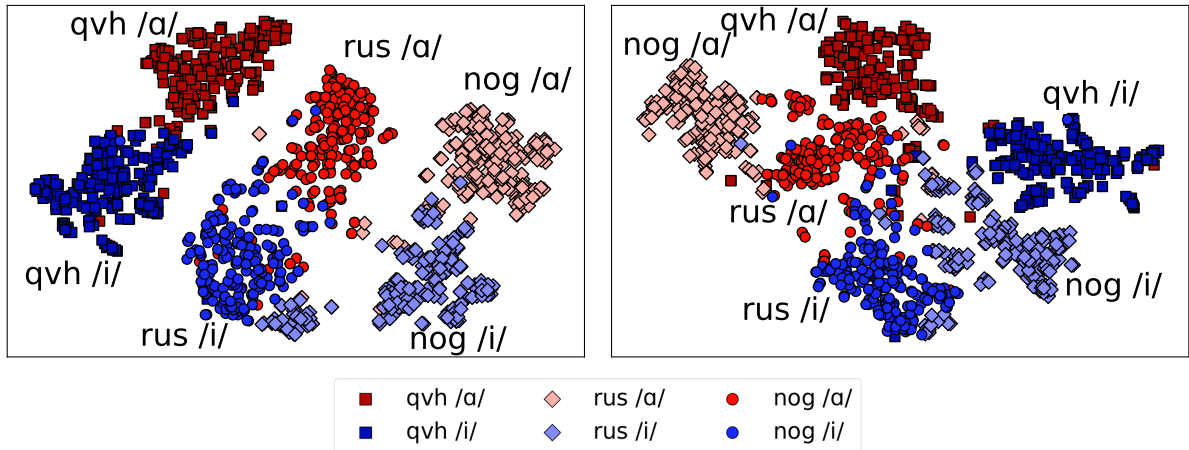
Figure 2: t-SNE representation of encoder states corresponding to /ɑ/ and /i/ across Quechua (Huamalies Dos de Mayo; qvh), Russian (rus), and Nogai (nog). *Left*: the model without the phoneme and adversarial objective. *Right*: the phoneme and language-adversarial objectives are added in, causing phoneme clusters between languages to gather closer together, and language to become less relevant in cluster placement.

amount of data in very closely related languages is available (in this case, close to 100 hours of QUE data), then there is little to be gained from highly unrelated languages.

When pretraining on QUE and CYR separately, the auxiliary objectives underperformed baseline multilingual pretraining on average. The variation in languages within these groups is far less than the variation between groups. Given that the phoneme and adversarial objectives are intended to overcome variation between pretraining languages, this result indicates that there must be a sufficient level of diversity in the pretraining languages before the auxiliary objectives are of benefit when adapting to certain target languages.

Results from pretraining on QUE+CYR showed either objective to help on average, and that the effects are complementary. Because of this, we opted to include them together in subsequent experimentation. We evaluated this best performing model on the larger set of other evaluation languages. Results in Table 3 show that in all cases multilingual pretraining of QUE+CYR with the auxiliary objectives outperformed its counterpart without the objectives (which frequently undeperformed the monolingual model), and in all but one case this led to an improvement over the monolingual baseline.[7]

To gain insight into how the auxiliary objectives change the representation of speech learnt by the

models, we applied 2D t-SNE dimensionality reduction (Van Der Maaten and Hinton, 2008). Figure 2 plots the representations of two phonemes in three languages learnt by the encoder[8] in the case without and with the auxiliary objectives. In the multilingual pretraining baseline, six clusters are represented for each language–phoneme combination. These appear stratified by language, with different phoneme clusters within languages close to one another. With the auxiliary objectives, phoneme clusters between languages move closer to one another, while language identity becomes less relevant in determining which phoneme clusters neighbour one another. In the latter plot, the Nogai phonemes become separated by a Russian /ɑ/. This is particularly salient since the Nogai speaker was female, while the Russian speaker had a deep male voice.

## 7 Reading Adaptation

In the previous section we explored the use of two dissimilar groups of languages in a multilingual setup. Multilingual pretraining of languages from a different language family and script benefitted from an explicit phoneme objective and adversarial objective when there was sufficient diversity in the pretraining languages. However, a change in orthography was conflated with a change in language family, geographic location, and phono-

---

[7]However, this doesn't hold in the *language adaptation* scenario, where the auxiliary objectives help QUE+CYR only slightly; see Section 8.

[8]We established the correspondence between encoder states and phonemes by using forced alignment with Kaldi (Povey et al., 2011), taking the encoder state at the mid-point of the duration on the phonemes.

| | MONO | QUE+CYR | | PHONOLOGY | | GEO | | 100-LANG | |
|---|---|---|---|---|---|---|---|---|---|
| | | - | +phn+adv | - | +phn+adv | - | +phn+adv | - | +phn+adv |
| ayr | 40.6 | 34.6 | 34.2 (-1.2%) | **33.9** | 34.5 (+1.8%) | 35.4 | 34.9 (-1.4%) | 34.2 | 34.5 (+0.9%) |
| quh | 14.8 | 14.9 | **13.9** (-6.7%) | 14.4 | 14.5 (+0.7%) | 15.5 | 14.8 (-4.5%) | 15.1 | 14.7 (-2.6%) |
| kek | 23.9 | 24.8 | 23.7 (-4.4%) | 24.8 | 24.5 (-1.2%) | 23.0 | **22.9** (-0.4%) | 24.9 | 24.4 (-2.0%) |
| ixl | 20.7 | 21.2 | 20.1 (-5.2%) | - | - | **19.7** | 20.1 (+2.0%) | 20.8 | 20.6 (-1.0%) |
| mlg | 45.2 | 43.5 | **41.4** (-4.8%) | 43.2 | 41.7 (-3.5%) | 43.3 | 42.2 (-2.5%) | 44.4 | 42.2 (-5.0%) |
| ind | 14.9 | 15.8 | 14.7 (-7.0%) | **13.7** | 14.3 (+4.4%) | 14.0 | **13.7** (-2.1%) | 14.7 | 14.2 (-3.4%) |
| kia | 14.6 | 14.6 | 13.2 (-9.6%) | - | - | **12.1** | 12.1 (-0.0%) | 14.4 | 13.0 (-9.7%) |
| swe | **20.5** | 22.7 | 21.6 (-4.9%) | 26.4 | 24.2 (-8.3%) | 22.0 | 21.2 (-3.6%) | 23.9 | 24.6 (+2.9%) |
| spn | 14.5 | 19.7 | 14.4 (-26.9%) | 13.9 | 13.8 (-0.7%) | 13.1 | **12.1** (-7.6%) | 15.8 | 14.8 (-6.3%) |
| | | Avg. rel. Δ: | (-7.8%) | Avg. rel. Δ: | (-1.0%) | Avg. rel. Δ: | (-2.3%) | Avg. rel. Δ: | (-2.9%) |

Table 3: Word error rate (%) comparison of adaptation of models pretrained on: Quechuan and Cyrillic-script languages (QUE+CYR), languages phonologically and phonetically similar to the target (PHON/INV), geographically proximate languages (GEO), and a massively multilingual set of languages (100-LANG). In each case we compared the average relative WER change when adding auxiliary phoneme and language-adversarial objectives (+phn+adv). Dashed entries had phonology and phonetic inventories that weren't well attested in URIEL, so were not assessed.

logical/phonetic characteristics. In this section, we investigate which factors are most important in choosing languages for multilingual pretraining and how useful it is to scale up model pretraining to many languages. This exploration is conducted in the reading adaptation scenario; language adaptation with unseen target speakers is addressed in Section 8. Beyond answering these questions, this investigation reveals more information about the utility of the proposed auxiliary objectives in different scenarios.

**Phonology & Geography** We test across a number of evaluation languages (c.f. Table 1) by determining, for each evaluation language, groups of pretraining languages that are similar to the evaluation languages in different ways. In order to determine language similarity in a principled way we used URIEL and *lang2vec* (Littell et al., 2017) to produce feature vectors for each language based on information from several linguistic resources before calculating their cosine similarity. For each language we used two feature vectors. The first is a concatenation of the lang2vec `phonology_average` and `inventory_average` vectors, characterizing phonological properties and phonetic inventory. The second represents geographic location. We denote these two groups PHON/INV and GEO respectively.[9] Geographic proximity may

serve as a proxy for other similarities not captured in PHON/INV, including language family, orthographic similarity, and the likelihood of exchanged loan words.

We filtered for languages in the dataset with `good` or `very good` alignments before ranking them by cosine similarity with the evaluation languages in terms of phonological and phonetic similarity as well as geographical proximity. To create each of the pretraining sets, we took between 7 and 14 of the top languages, matching approximately the total duration of the phonetically/phonologically similar groups with the geographically proximate language groups.[10] For most languages, there is no overlap between the GEO and PHON/INV sets.

**Massively multilingual model** As a further point of comparison, we pretrain a model on around 100 languages (denoted 100-LANG), for approximately 1650 training hours in total.[11]

### 7.1 Auxiliary Objectives Findings

The results in Table 3 extend on our findings in Section 6, continuing to support the benefit of the use of the auxiliary objectives while shedding more light on the type of language variability the objectives help to overcome. GEO and 100-LANG

---

[9] We didn't create PHON/INV sets for Ixil and Garap because their phonological features and phonetic inventories were not well attested, and we didn't use the lang2vec language family vectors since most of the Quechuan languages were not captured as being highly similar to SB Quechua.

[10] An exhaustive list of the CMU Wilderness language codes for each pretraining group can be found in Appendix A, along with durations of each pretraining set.

[11] These models were pretrained for 6 epochs.

benefitted comparably from the objectives on average, while PHON/INV did less so. QUE+CYR benefitted the most. This suggests that the objectives may help more when pretraining languages are orthographically, phonetically and phonologically diverse.

Unlike the other languages, the Swedish PHON/INV vectors were not well attested. As a result the Swedish PHON/INV group has languages with a similar phonetic inventory that were also unattested phonologically. This model underperformed the monolingual model by a large margin, suggesting that similarity of phonetic inventory alone may not be so useful alone without similarity of phonological features. Models pretrained on this set also benefitted the most from the auxiliary objectives. It may be the case that the auxiliary objectives push together representations of allophones within languages, and pronunciation variations of the same phonemes between languages. When Swedish is discounted, the average relative improvement when adding auxiliary objectives for PHON/INV becomes negligable.

The PHON/INV configurations are hurt by the auxiliary objectives for SB Quechua and Aymara and Indonesian. The PHON/INV sets for the first two of these languages emphasized Quechuan languages, and this corroborates the indication in Section 6 that the auxiliary objectives may not help so much when pretraining languages are similar. On the other hand, the Indonesian PHON/INV included Afro-Asiatic and Niger-Congo languages, as well an Indo-European language and Huave, a language isolate from Mexico, yet it was not improved by auxiliary objectives.

## 7.2 Choice of Pretraining Languages

The average relative word error rate (WER) change for GEO against PHON/INV was -2.2% without auxiliary objectives, and -4.4% with them,[12] suggesting that features correlated with geography are useful for guiding pretraining language selection. Counter-examples were Aymara, SB Quechua and Malagasy, which performed worse when pretrained on GEO. In the case of SB Quechua, only one Quechuan language was represented in GEO (Inga), while PHON/INV had three (qub, qvh, quf). Madagascar is far removed from where most Austronesian languages are spoken, so Malagasy's GEO set were almost all Niger-Congo
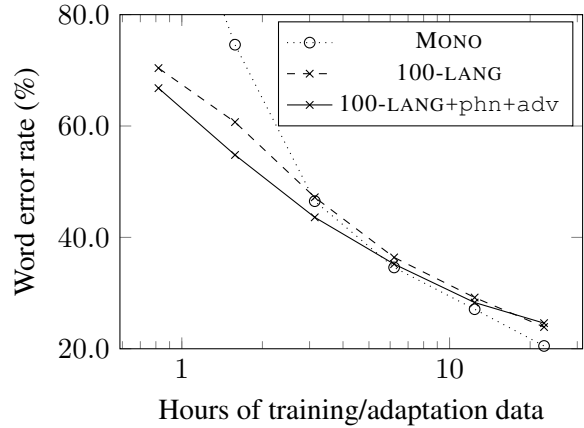


Figure 3: Scaling training/adaptation data for Swedish. Adapting to the full dataset, the auxiliary objectives underperformed both the monolingual and baselines, but yields an advantage when the model is adapted to less target language data.

languages, while the PHON/INV had a diverse array of Austronesian, Indo European, Afro-Asiatic, Sino-Tibetan and Mayan languages. However, on average, these results suggest that geographical proximity is a decent guide to pretraining language selection. Another advantage is that it requires no explicit phonological features, making it applicable to a much larger number of languages.

The average relative WER change of 100-LANG against MONO was +1.3%, indicating that massively multilingual pretraining by itself not useful if the target speakers are seen in training. Using the auxiliary objectives overcame the difference, resulting in a -1.6% average relative WER change. However, pretraining with GEO+phn+adv yielded an average relative delta of -7.4% over the monolingual model. Though more languages help, they are not necessarily better than geographically proximal languages (however, results are very different when not adapting to target speakers: see Section 8).

In two cases pretraining with 100-LANG was hindered by the auxiliary objective. In one of these cases, Swedish, both 100-LANG variations substantially underperformed the monolingual baseline. One possible reason is that there is enough target language and speaker data that the multilingual pretraining and auxiliary objectives offer no benefit. We scaled training/adaptation data for Swedish from under 1 hour. Figure 3 indicates that in this case the auxiliary objectives do lead to better initialization, with gains being lost only when around 5 hours of target language and reading data

---

[12]Discounting Swedish, this becomes +0.2% and -3.1%.

are seen.

## 8 Language Adaptation

Previous sections have addressed the reading adaptation scenario, where the ASR model is adapted to speech from the target reading (ie. where target speakers have been heard in adaptation). In this section we evaluate in a language adaptation scenario, adapting to readings in the target language, but not the target reading. The question of how well a multilingual model can be adapted to a language on the basis of recordings from a small number of target-language speakers is relevant to incident response situations such as those modelled by LORELEI (Strassel and Tracey, 2016), where a single language consultant is available for which recorded speech can be made. We performed experiments analogous to those of the previous sections where the evaluation reading was not seen in training or adaptation. This is a challenging task as the model must generalize to multiple speakers of a language on the basis of seeing only several in training. Most of the findings corroborate what was found in the previous sections. Here we highlight differences.

**Massively multilingual pretraining** led to substantially better performance than other methods, unlike in the reading adaptation task. For each evaluation language, the 100-LANG model outperformed the next best method, with one exception: Indonesian. In that case GEO set performed the best, as the languages were not only geographically proximate, but also consisted entirely of other Austronesian languages. The takeaway (c.f. Table 4) is that you should always use more pretraining languages unless you know your target speakers, as in the reading adaptation scenario.

**Auxiliary objectives** remained useful on the whole. However, while the difference in WER achieved when adding the auxiliary objectives was similar to those reported in Section 7 for PHON/INV and 100-LANG, GEO and QUE+CYR no longer achieved improvements. QUE+CYR notably only achieved a -0.2% average relative WER change when adding the auxiliary objectives, while achieving -7.8% in the reading adaptation case. While the auxiliary objectives remained useful on the whole, their effect was dwarfed by the value of adding more languages.

| | MONO | QUE +CYR | PHON +INV | GEO | 100-LANG |
|---|---|---|---|---|---|
| ayr | 91.4 | 86.3 | 86.7 | 87.2 | **79.2** (-8.2%) |
| quh | 62.3 | 35.8 | 35.5 | 42.8 | **30.1** (-15.2%) |
| kek | 75.6 | 74.3 | 73.8 | 74.4 | **73.5** (-0.4%) |
| ixl | 81.8 | 79.8 | - | 78.4 | **74.3** (-6.9%) |
| mlg | 103.6 | 68.3 | 64.0 | 63.7 | **62.2** (-2.4%) |
| ind | 24.6 | 23.5 | 22.1 | **21.1** | 21.6 (+2.4%) |
| kia | 57.2 | 51.5 | - | 49.9 | **48.2** (-6.4%) |
| swe | 72.9 | 64.4 | 75.4 | 62.5 | **55.1** (-11.8%) |
| spn | 44.8 | 33.8 | 33.4 | 32.7 | **29.9** (-8.6%) |
| Avg. rel. Δ of 100-LANG wrt. next best method: | | | | | (-6.0%) |

Table 4: Adaptation to the non-target reading in the target language. All language sets use the auxiliary training objectives, which again exhibited an relative gain over the corresponding model without. The relative deltas of 100-LANG are with respect to the next closest model on a language-by-language basis.

**Phonology versus Geography** GEO sets with or without auxiliary objectives lost their edge over PHON/INV, with high variance in scores. The amount of training data becomes the dominating variable affecting WER.

## 9 Conclusions

We have explored the utility of pretraining multilingual models on a variety of language sets, scaling to as as many as 100 languages. Our experiments have demonstrated the value of auxiliary phoneme and language-adversarial pretraining objectives in a multilingual end-to-end ASR framework, particularly when the pretraining languages are diverse. Our results suggest how to pick pretraining languages when target speakers are seen in the adaptation data: find geographically proximal languages. When adapting to just several non-target speakers, exposure to more speech in pretraining is the most important thing for model generality, even if from a wide range of dissimilar languages.

## Acknowledgments

## References

Alan W Black. 2019. CMU Wilderness Multilingual Speech Dataset. In *ICASSP*.

William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. 2015. Listen, attend and spell. In *ICASSP*.

Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Katya Gonina, et al. 2018. State-of-the-art speech recognition with sequence-to-sequence models. In *ICASSP*.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. *Advances in Neural Information Processing Systems 28*, pages 577–585.

Siddharth Dalmia, Ramon Sanabria, Florian Metze, and Alan W Black. 2018. Sequence-based multilingual low resource speech recognition. In *ICASSP*.

Van Hai Do, Xiong Xiao, Eng Siong Chng, and Haizhou Li. 2014. Cross-lingual phone mapping for large vocabulary speech recognition of under-resourced languages. *IEICE Transactions on Information and Systems*, E97-D(2).

Mark J F Gales, Kate M Knill, Anton Ragni, and Shakti P Rath. 2014. Speech recognition and keyword spotting for low-resource languages: BABEL project research at CUED. In *Spoken Language Technologies for Under-Resourced Languages*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1).

A Graves, A.-R. Mohamed, and G Hinton. 2013. Speech recognition with deep recurrent neural networks. *ICASSP*.

Alex Graves, Santiago Fernandez, Faustino Gomez, and Jurgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. *ICML*.

Georg Heigold, Vincent Vanhoucke, Alan Senior, Patrick Nguyen, Marc'Aurelio Ranzato, Matthieu Devin, and Jeffrey Dean. 2013. Multilingual acoustic models using distributed deep neural networks. In *ICASSP*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8).

Takaaki Hori, Shinji Watanabe, Yu Zhang, and William Chan. 2017. Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM. In *INTERSPEECH*.

David Imseng, Petr Motlicek, Hervé Bourlard, and Philip N Garner. 2014. Using out-of-language data to improve an under-resourced speech recognizer. *Speech Communication*, 56.

Hirofumi Inaguma, Jaejin Cho, Murali Karthick Baskar, Tatsuya Kawahara, and Shinji Watanabe. 2018. Transfer learning of language-independent end-to-end ASR with language model fusion. *arXiv:1811.02134*.

Martin Karafiát, Murali Karthick Baskar, Shinji Watanabe, Takaaki Hori, Matthew Wiesner, and Jan "Honza" Černocký. 2018. Analysis of multilingual sequence-to-sequence speech recognition systems. *arXiv:1811.03451*.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2016. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In *ICASSP*.

Kalpesh Krishna, Shubham Toshniwal, and Karen Livescu. 2018. Hierarchical multitask learning for CTC-based speech recognition. *arXiv:1807.06234*.

Viet Bac Le and Laurent Besacier. 2005. First steps in fast acoustic modeling for a new target language: application to Vietnamese. In *ICASSP*.

Patrick Littell, David Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *EACL*.

Alexis Michaud, Oliver Adams, Trevor Anthony Cohn, Graham Neubig, and Séverine Guillaume. 2018. Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit. *Language Documentation & Conservation*, 12.

Markus Müller, Sebastian Stüker, and Alex Waibel. 2017. Phonemic and Graphemic Multilingual CTC Based Speech Recognition. *arXiv:1711.04564*.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. *NeurIPS*.

Christian Plahl, Ralf Schlüter, and Hermann Ney. 2011. Cross-lingual portability of Chinese and English neural network features for French and German LVCSR. In *Automatic Speech Recognition and Understanding (ASRU), IEEE Workshop on*.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, and Others. 2011. The Kaldi speech recognition toolkit. In *Automatic Speech Recognition and Understanding (ASRU), IEEE workshop on*.

Kanishka Rao and Haim Sak. 2017. Multi-accent speech recognition with hierarchical grapheme based models. *ICASSP*.

Ramon Sanabria and Florian Metze. 2018. Hierarchical multi task learning with CTC. In *IEEE Spoken Language Technology Workshop (SLT)*.

Stefano Scanzio, Pietro Laface, Luciano Fissore, Roberto Gemello, and Franco Mana. 2008. On the use of a multilingual neural network front-end. In *INTERSPEECH*.

Odette Scharenborg, Francesco Ciannella, Shruti Palaskar, Alan Black, Florian Metze, Lucas Ondel, and Mark Hasegawa-Johnson. 2017. Building an ASR system for a low-research language through the adaptation of a high-resource language ASR system: preliminary results. In *International Conference on Natural Language, Signal and Speech Processing (ICNLSSP)*.

Tanja Schultz. 2002. GlobalPhone: a multilingual speech and text database developed at Karlsruhe University. In *Seventh International Conference on Spoken Language Processing*.

Tanja Schultz and Alex Waibel. 2001. Experiments on cross-language acoustic Modeling. *EUROSPEECH'01*.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11).

Tom Sercu, Christian Puhrsch, Brian Kingsbury, and Yorktown Heights. 2016. Very deep multilingual convolutional neural networks for LVCSR. In *ICASSP*.

Yusuke Shinohara. 2016. Adversarial multi-task learning of deep neural networks for robust speech recognition. In *INTERSPEECH*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*.

A. Stolcke, F. Grezl, Mei-Yuh Hwang, Xin Lei, N. Morgan, and D. Vergyri. 2006. Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons. In *ICASSP*.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI language packs: data, tools, and resources for technology development in low resource languages. *LREC*.

Sining Sun. 2018. Domain adversarial training for accented speech recognition. In *ICASSP*.

Paul Taylor, Alan W Black, and Richard Caley. 1998. The architecture of the Festival speech synthesis system. In *ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*.

Samuel Thomas, Sriram Ganapathy, Hynek Hermansky, and Speech Processing. 2012. Multilingual MLP features for low-resource LVCSR systems. In *ICASSP*.

Shubham Toshniwal, Tara N. Sainath, Ron J. Weiss, Bo Li, Pedro Moreno, Eugene Weinstein, and Kanishka Rao. 2017. Multilingual speech recognition with a single end-to-end model. In *ICASSP*.

László Tóth, Joe Frankel, Gábor Gosztolya, and Simon King. 2008. Cross-lingual portability of MLP-based tandem features - a case study for English and Hungarian. *INTERSPEECH*.

Aditay Tripathi, Aanchan Mohan, Saket Anand, and Maneesh Singh. 2018. Adversarial learning of raw speech features for domain invariant speech recognition. In *ICASSP*.

Laurens Van Der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. Technical report.

Karel Vesely, Martin Karafiát, Frantisek Grezl, Marcel Janda, and Ekaterina Egorova. 2012. The language-independent bottleneck features. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*.

Ngoc Thang Vu, Florian Metze, and Tanja Schultz. 2012. Multilingual bottle-neck features and its application for under-resourced languages. In *The third International Workshop on Spoken Language Technologies for Under-resourced languages*.

Shinji Watanabe, Takaaki Hori, and John R Hershey. 2017a. Language independent end-to-end architecture for joint language identification and speech recognition. In *Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE Workshop on*.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. 2017b. Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal on Selected Topics in Signal Processing*.

Jiangyan Yi, Jianhua Tao, Zhengqi Wen, and Ye Bai. 2018. Adversarial multilingual training for low-resource speech recognition. *ICASSP*.

Su-Youn Yoon, Kyoung-Young Kim, and Richard Sproat. 2007. Multilingual transliteration using feature based phonetic method. In *ACL*.

# A List of readings in each language set

Below is a collection of lists of the CMU Wilderness reading codes that comprise different groupings. This includes the target language readings; the Quechuan group; the Cyrillic-script group; the phonologically similar and geographically similar sets for each target language; and the massively multilingual set.

**Target language readings** MLGEIV, ML-GRCV, MLGRPV, IX1WBT, IXIWBT, IXLWBT, INZNTV, INZSHL, INZTSI, QUHRBV, QUHSBB, QEJLLB, QUBPBS, QUFLLB, QVSTBL, QVWTBL, QWHLLB, SPNBDA, SPNWTC, KIABSC, KIAWBT, KEKIBS, KEKSBG, SWESFB, SWESFV, AYMSBU, AYMBSB.

**Evaluation readings** AYMSBU, MLGRPV, IXIWBT, INZSHL, QUHRBV, SPNBDA, KIAWBT, KEKIBS, SWESFV.

**QUE (97.6 training hours)** QEJLLB, QUBPBS, QUFLLB, QVSTBL, QVWTBL, QWHLLB.

**CYR (59.6 training hours)** NOGIBT, BAKIBT, GAGIB1, KJHIBT, RUSS76, CRHIBT.

**AYR-PHON/INV (145.3 training hours)** QUBPBS, TOBBSA, QUFLLB, QVSTBL, INBWBT, QEJLLB, JICWBT, QU1LSM, QUTIBS.

**AYR-GEO (146.2 training hours)** IGNSBB, TNATBL, GNWNTM, ESENTM, MCBTBL, GYRSBB, CBSBSP

**QUH-PHON/INV (177.9 training hours)** TOBBSA, DUGBTL, QUBPBS, TZHSBM, HUSLLB, NYFBTL, NCUWBT, QEJLLB, QUFLLB, HAGGIL, NZIBSG, MNBTBL.

**QUH-GEO (178.5 training hours)** GNWNTM, IGNSBB, TOBBSA, ENXBSP, GYRSBB, CAXSBB, CEGNTP, TNATBL, ESENTM, TERTBL.

**KEK-PHON+INV (142.1 training hours)** QU1LSM, QUTIBS, TZTWBT, TUFWYI, QWHLLB, PAGPBS, UDUSIM, YUASBM.

**KEK-GEO (137.0 training hours)** MOPWBT, POHBSG, CA1WBT, CKIWBT, TZTWBT, QU1LSM, QUTIBS, BZJBSW.

**MLG-PHON/INV (198.2 training hours)** RONBSR, TGLPBS, KVNWBT, HUVTBL, KBRSIM, TPMWBT, BTXLAI, KACUBS, WMWWYI, IGNSBB, HAEBSE, IBATIV, HILHPV, TZBSBM.

**MLG-GEO (205.38 training hours)** WMWWYI, VMWBSM, MFEBSM, SEHBSM, TOHSBM, CCESBM, KDCPBT, CWEPBT, KKIBST, NYYBST, KSBBST, KDNBSZ, DUGBTL, GOGBST.

**IND-PHON/INV (193.1 training hours)** IBATIV, TGLPBS, HAEBSE, KERABT, KACUBS, NYFBTL, RONBSR, CWTATB, HUVTBL, BTXLAI, IGNSBB, JAVNRF, DUGBTL, MNKBSG.

**IND-GEO (191.5 training hours)** SUNIBS, NIJLAI, JAVNRF, PSELAI, IBATIV, PTULAI, MVPLAI, PPKLAI, BEPLAI, NPYLAI, LEWLAI, MWVLAI.

**SWE-PHON/INV (122.4 training hours)** KDJBSU, NZIBSG, ANVWBT, DGABSG, SHKBSS, SLDTBL, KUSTBL, MUYWBT, NCUWBT, LIABSL, CKOGIL.

**SWE-GEO (122.4 training hours)** RMCWFW, EN1NIV, RMORAM, RONBSR, GAGIB1, GAGIBT, CRHIBT, KPVIBT, LTNNVV, ALSBSA, UDMIBT, XALIBT, BAKIBT.

**SPN-PHON/INV (123.7 training hours)** KVNWBT, HAEBSE, HUVTBL, GUGRPV, HUSLLB, GUMTBL, NYFBTL, KWIWBT.

**SPN-GEO (129.5 training hours)** PORARA, LTNNVV, EN1NIV, RMORAM, ALSBSA, RMCWFW, RONBSR, GAGIB1, GAGIBT, CRHIBT, TAQWBT, FUQWBT, MYKWBT.

**100-LANG (1646.8 training hours)** OBOWBT, ACUTBL, SEYWBT, HAUCLV, BZHPNG, AMKWBT, GAGIB1, GNWNTM, URBWBT, RUGWBT, PAUUBS, SEHBSM, SNNWBT, KQETBL, TGOTBL, NOGIBT, XTMTBL, OJ1CBS, TNATBL, AIAWYI, PABTBL, MEJTBL, TWBOMF, HUSLLB, ESENTM, BAKIBT, HNNOMF, IFAWBT, ENXBSP, ALJOMF, PXMBSM, JAISBG, PIRWBT, DOMBEC, NINWYI, BEPLAI, JAMBSW, TERTBL, LAWNTM, URATBL, AGNWPS, TPIPNG, TTCWBT, HUUTBL, NPYLAI, KJHIBT, AZZTBL, COKWBT, KWIWBT, SABWBT, PADTBL, GUMTBL, CRHIBT, QXRBSE, RMORAM, NHYTBL, TPPTBL, TUFWYI, ZLMAVB, PRFWBT, TWULAI, GAGIBT, FARWBT, OM1TBL, RUSS76, PTULAI, MIFWBT, MIYWYI, MRWNVS, KNETBL, PBCBSS, MYYWBT, ACHBSU, ACNBSM, ADETBL, AHKTBS, AK1BSG, ALPWBT, ALSBSA, ALTIBT, ANVWBT, ATGWYI, AVNWBT, AVUWBT, AYMBSB, AYMSBU, AZEBSA,

BEXWBT, BQJATB, BTXLAI, BZJBSW, CA1WBT, CARBSS, CAXSBB, CBSBSP, CMRWBT, CNLTBL, CNMRGB, CRNWBT.