Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6420–6428
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Progress in Multilingual Speech Recognition for Low Resource Languages Kurmanji Kurdish, Cree and Inuktut

Vishwa Gupta and Gilles Boulianne
Centre de Recherche Informatique de Montréal (CRIM), Quebec, Canada
{vishwa.gupta,gilles.boulianne}@crim.ca

Abstract

This contribution presents our efforts to develop the automatic speech recognition (ASR) systems for three low resource languages: Kurmanji Kurdish, Cree and Inuktut. As a first step, we generate multilingual models from acoustic training data from 12 different languages in the hybrid DNN/HMM framework. We explore different strategies for combining the phones from different languages: either keep the phone labels separate for each language or merge the common phones. For Kurmanji Kurdish and Inuktut, keeping the phones separate gives much lower word error rate (WER), while merging phones gives lower WER for Cree. These WER are lower than training the acoustic models separately for each language. We also compare two different DNN architectures: factored time delay neural network (TDNN-F), and bidirectional long short-term memory (BLSTM) acoustic models. The TDNN-F acoustic models give significantly lower WER for Kurmanji Kurdish and Cree, while BLSTM acoustic models give significantly lower WER for Inuktut. We also show that for each language, training multilingual acoustic models by one more epoch with acoustic data from that language reduces the WER significantly. We also added 512-dimensional embedding features from cross-lingual pre-trained wav2vec2.0 XLSR-53 models, but they lead to only a small reduction in WER.

Keywords: multilingual acoustic models, OpenASR20 Challenge, factored TDNN, bidirectional LSTM, wav2vec embedding

1. Introduction

The low resource transcription and keyword spotting effort received a great impetus from the IARPA Babel program[1]. The main goal of the program was to improve the performance of keyword search on languages with very little transcribed data (low-resource languages). Data from 26 languages was collected with certain languages being held out as *surprise* languages to test the ability of the teams to rapidly build a system for a new language[2]. Many different DNN training algorithms have been experimented with within the Babel program (Gales et al., 2014) (Knill et al., 2014) (Huang et al., 2013) (Trmal et al., 2014) (Chen et al., 2013) (Zhang et al., 2014). In (Gales et al., 2014) they experiment with both DNN and tandem systems and achieve token error rates (TER) between 60% and 77% with limited language packs (10 hours of training audio), depending on the language and training algorithms. They also experiment with data augmentation by automatically labeling untranscribed data.

In (Huang et al., 2013), the authors keep the senones separate for each language in the softmax layer, while sharing the hidden layers across language. In (Vu et al., 2014), the authors experiment with two different ways of generating multilingual phone sets: keep phones for each language separate, or merge phones that have the same IPA symbols across different languages. In their study, they show that keeping phones separate for each language in multilingual DNN training results in lower word error rate (WER) compared to merging phones across languages for small amount of training data (1 hour). For larger acoustic training data per language, the two multilingual DNNs (with and without merging) give similar WER.

In (Li et al., 2020), the authors experiment with three different ways of modeling multilingual phone sets in order to generate good universal phone set with good coverage across many low resource languages: private (separate phones for each language), shared (pooling phones with the same IPA symbol across languages), and allosaurus (predict over a shared phone inventory, then map into language-specific phonemes with an allophone layer). They show that for phone recognition, private phone set gives significantly lower phone error rate (PER) than the shared phone set, while their allosaurus system improves on the separate phone set.

In E2E (end-to-end) multilingual ASR, if the text scripts use different character sets for each language, then there is a natural separation of end-to-end modeling for each language. For example, in (Dalmia et al., 2018), the authors train E2E models to improve speech recognition for low resource Babel languages. In (Kannan et al., 2019), the authors train E2E models for 9 Indian languages. All the languages except two have different scripts which separates the language training. However, there is significant overlap of vocabulary in these Indian languages, and many words are written in Latin that result in higher error rates for languages with less training data. So they condition the RNN-T encoder on a language vector.

[1] https://www.iarpa.gov/index.php/research-programs/babel

[2] https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/harper.pdf

In (Conneau et al., 2020), the authors present a cross-lingual speech representation (XLSR) system which learns cross-lingual speech representations by pretraining a single model from the raw waveform of speech in multiple languages. On the Common Voice benchmark, XLSR leads to significant reduction in phoneme error rate. On BABEL also, the XLSR system shows significant reduction in word error rate. So in one experiment, we also computed 512-dimensional embeddings from this pretrained XLSR-53 system to see if we can reduce word error rates even further for our multilingual system.

In this paper, we experiment with training acoustic models with different multilingual phone sets from 12 different low resource languages to minimize word error rate (WER). For each language, we have anywhere from 20 hours to 80 hours for training the multilingual models. For decoding, we decode test sets for 3 morphologically complex languages for which we have more than 40 hours of audio for training: Kurmanji Kurdish, Cree and Inuktut. In this scenario, we find that multilingual acoustic models with separate phone sets per language give significantly lower WER (than unilingual acoustic models for each language) for two of the languages. For Cree, the multilingual acoustic model trained with merged phone set (from the 12 languages) gives lower WER. This is in contrast to the results in (Vu et al., 2014) where multilingual acoustic models with separate phone sets gave significantly lower WER only for small training sets (1 hour of audio for each language).

Another issue is whether different DNN architectures make a difference in multilingual acoustic modeling. So we experimented with two different architectures: factored time delay neural networks (TDNN-F) (Povey et al., 2018) and bidirectional LSTM (BLSTM) (Graves et al., 2013) acoustic models used widely in the literature. For Kurmanji Kurdish and for Cree, TDNN-F acoustic models significantly outperformed the BLSTM acoustic models, while for Inuktut, BLSTM acoustic models gave significantly lower WER than the TDNN-F acoustic models. We note that Inuktut is a highly polysynthetic language (Schwartz et al., 2020) (Gupta and Boulianne, 2020a), where we decode syllables and then merge syllable sequences into word sequences. For Kurmanji Kurdish and for Cree we decode word sequences since we can have a small dictionary with low out-of-vocabulary (OOV) rate for unseen text, even though Kurmanji Kurdish is an agglutinative language and Cree is polysynthetic (although much less than Inuktut).

We also found that training the multilingual models for one more epoch with a small learning rate from acoustic training data from one language results in significant reduction in WER for that language.

Another issue is whether we can add features from pretrained cross-lingual XLSR models (Conneau et al., 2020) to reduce WER for multilingual speech recognition. So we generated 512-dimensional embeddings (cross-lingual speech representations) using the publicly available pretrained XLSR-53 model and added them to 40-dimensional MFCC features. The enhanced features reduced the WER for two development sets by a small amount, so we probably need to adapt the pretrained XLSR model with our multilingual acoustic data to get a bigger impact on WER.

In summary, we revisit strategies and architectures for training multilingual acoustic models but, differently from previous studies, we include several morphologically complex languages in our test. We present four main contributions in this paper. First, for separate vs shared phoneme sets, we find that separate sets perform better, in contrast to previous studies. Second, we observe that a TDNN-F architecture outperforms BLSTM, with the exception of the highly polysynthetic language. Third, we find that incremental fine-tuning a multilingual model on the target language leads to significant reduction in WER for that language. Finally, we report experiments with pre-trained multilingual acoustic embeddings.

2. Data Resources

For multilingual acoustic model training, we used training data provided by NIST for OpenASR20 challenge[3] for 10 low resource languages: Amharic, Cantonese, Guarani, Javanese, Kurmanji Kurdish, Mongolian, Pashto, Somali, Tamil, and Vietnamese. For each of these languages there is close to 20 hours of audio. We added acoustic data we have for two Canadian indigenous languages: Inuktut and Cree. Our focus in this paper is on 3 languages: Kurmanji Kurdish, Cree and Inuktut. So we added more acoustic training data for these 3 languages. Table 1 summarizes the various datasets we used for acoustic and language model training for these three languages. For Kurmanji Kurdish, we added data from the LDC Kurmanji Kurdish corpus[4] and also data we received from TWB (Translators without Borders). The openASR20 training data for Kurmanji Kurdish was part of the data from LDC Kurmanji Kurdish corpus. This LDC corpus contains 153 hours of transcribed audio for training. After removing the noise and silence portion, we have 51 hours of transcribed audio. The LDC disk also contained 88 hours of untranscribed audio. After voice activity detection and automated transcription of the untranscribed audio, we added 19.35 hours of reliable segments for training. We followed a process similar to that for Cree (Gupta and Boulianne, 2020b) to extract these reliable segments. Our best acoustic model for Kurmanji Kurdish was trained from 70.4 hours of audio.

For Cree, we used 30 hours of transcribed audio and 71 hours of automatically transcribed audio from approximately 1045 hours of untranscribed CBC radio record-

[3] https://www.nist.gov/itl/iad/mig/openasr-challenge

[4] IARPA Babel Kurmanji Kurdish Language Pack, IARPA-babel205b-v1.0a (LDC2017S22)

ings (Gupta and Boulianne, 2020b). The training audio for Cree is much larger than in (Gupta and Boulianne, 2020b) and the development set has changed to 6 hours of CBC Cree recordings. For Inuktut, we used roughly 78 hours of transcribed stories for training, larger than in (Gupta and Boulianne, 2020a), but we have kept the development set to be the same (3 hours of audio) as in (Gupta and Boulianne, 2020a).

The additional transcribed audio for Inuktut and Cree came from the National Research Council of Canada (NRC) funded project on Indigenous languages (Kuhn et al., 2020). This NRC project developed software to assist Indigenous communities in Canada in preserving their languages and extending their use. Through this NRC project, audio from many Indigenous languages was transcribed for future use in developing speech-to-text transcription for those Indigenous languages also.

The OpenASR20 development set for Kurmanji Kurdish (see Table 2) was 17.8 hours of audio with transcripts. The transcribed audio for one speaker from TWB (containing 5047 sentences and 58808 words from a corpus of tales) was used as another Kurmanji Kurdish development set. The tales text was excluded from any language model training.

For Cree we used 6 transcribed CBC recordings for development (roughly 1 hour per recording), and for Inuktut 7 transcribed recordings (a total of 3 hours of audio) as shown in Table 2.

Source	Audio	Text
Cree (CBC radio recordings)	101.0 h	245.1 k
Inuktut (Nunavut)	78 hours	172.1 k
Kurmanji OpenASR20	18.0 h	81.2 k
Kurmanji LDC transcribed	51.0 h	272.1 k
Kurmanji LDC untranscribed	19.4 h	187.9 k
TWB-news	-	45.0 M
TWB-other	-	1.1 M
TWB-all	-	46.1 M

Table 1: *Training sets, in hours of transcribed audio and number of words in texts. TWB-all includes TWB-news and TWB-other.*

Source	Audio	Text
OpenASR20 (Kurmanji)	17.8 h	76.7 k
TWB-audio (Kurmanji)	6.5 h	58.8 k
Cree	6.0 h	31.3 k
Inuktut	3.0 h	7.7 k

Table 2: *Development sets (in hours and number of words).*

2.1. Language models

The training lexicons for the 10 languages in openASR20 were taken from the openASR20 build. For Kurmanji Kurdish, we used the lexicon from the LDC Kurmanji Kurdish corpus. For Inuktut and Cree, which have writing systems very close to phonetic, we used a simple set of rules derived from descriptions of the writing systems. All the lexicons used IPA X-SAMPA[5] phone symbols.

Language	Voc	TTR	OOV	PPL	Size
Cree	90k	0.26	10.3%	230	986k
Inuktut-w	129k	0.61	63.0%	1250	2.64M
Inuktut-s	3.2k	-	0.0%	30.8	1.44M
Kurmanji	6k	0.21	6.3%	133	600k

Table 3: *Language model details for each language. Inuktut-w is word-based while Inuktut-s is syllable-based. TTR is the ratio of types to tokens. OOV is out-of-vocabulary rate. PPL is perplexity on dev set. Size is the number of parameters in the model.*

Table 3 summarizes language model characteristics for each language, in terms of vocabulary, out-of-vocabulary rate and perplexity. We computed TTR, the ratio of the number of types (vocabulary size) to the number of tokens (word occurrences), by randomly selecting sentences from training texts until 50k tokens were collected, then counting types. TTR gives an indication of morphological complexity and is well correlated with other linguistic measures of complexity (Bentz et al., 2016) such as the degree of polysynthetism. With a TTR of 0.61, Inuktitut appears much more complex than Cree, which is also polysynthetic, and Kurmanji Kurdish, which is agglutinative. In comparison, English, on the same sample size, obtains a TTR of 0.16 on the LibriSpeech corpus and 0.11 on Switchboard.

For Kurmanji Kurdish language model training, we had a total of 46.44 million words of text: 81.2k words from OpenASR20 training text, 272k words of additional text from LDC, and 46.1 million words of text from TWB. TWB text can be divided into two parts: 45 million words from TWB-news[6] containing news (45.7% accented words), and TWB-other from other sources with 1.1 million words (46.9% accented words).

A separate trigram language model (LM) is trained for each development set. For the OpenASR20 dev set, the LM is trained from Kurmanji Kurdish LDC transcribed + untranscribed text (see Table 1). For TWB-audio, we train the LM from 46.1 million words of text from TWB. For *CBC Cree 6files* development set, we train the LM from all the Cree training text plus all the text in East Cree we could find over the internet (approximately 253 k words (see (Gupta and Boulianne, 2020b) for details), and for *Inuktut 7files* development set, we train a 4-gram syllabic LM from 53k words of Inuktut training text + Nunavut parliament proceedings (Hansard) containing 6.5 million words of text (see (Gupta and Boulianne, 2020a) for details). The Inuktut language model for decoding is syllabic with the

[5] https://fr.wikipedia.org/wiki/X-SAMPA

[6] https://anfkurdi.com/

dictionary containing 3158 context dependent (begin, middle or end of word) syllables. The decoded syllable sequences are transformed into word sequences with the help of syllables with word-end markers in the decoded syllable sequence.

3. Optimization of Multilingual Training

We trained multilingual DNNs from the acoustic training data from all the 12 languages as described in the *data resource* section. For testing, we used development sets from three languages: Kurmanji Kurdish, Cree and Inuktut. Kurmanji Kurdish is one of the ten languages in the OpenASR20 challenge. Inuktut and Cree are two of the most spoken indigenous languages in Canada (out of 70+ indigenous languages spoken in Canada), and we have significant amount of data for both these languages. Inuktut is a highly polysynthetic language (Schwartz et al., 2020) and new words are generated by concatenating morphemes to the existing word. One word in Inuktut can represent a whole phrase or sentence in English. That is why, even with a very large dictionary, the out-of-vocabulary rate for a new text can be 60% or higher. For this reason, we use syllabic decoding for Inuktut, and convert syllable sequences to word sequences by concatenating syllables (Gupta and Boulianne, 2020a). This syllable sequence to word sequence conversion is facilitated by adding word-end markers to syllables, and these decoded syllables with word-end markers are used to find word boundaries. So it will be interesting to see if such linguistic differences affect acoustic modeling.

To optimise acoustic modeling, we tried two different multilingual phone sets: one phone set where the non-silence phones for each language are kept separate by adding a language tag to each phone, and second phone set where all the common phones in the 12 languages are merged. We also tried two different DNN architectures: factored time delay neural network (TDNN-F) (Povey et al., 2018) and bidirectional LSTM neural network (BLSTM) (Graves et al., 2013). We also tried two different feature parameters: 40-dimensional MFCC features, and these MFCC features concatenated with 512-dimensional embeddings obtained from pretrained cross-lingual XLSR-53 wav2vec2.0 model (Conneau et al., 2020). In order to compare multilingual results with unilingual results, we first give word error rates (WER) when we train acoustic models separately for each language (unilingual results).

3.1. Separate Training for Each Language

For the three languages Kurmanji Kurdish, Cree and Inuktut, we trained factored TDNN (TDNN-F) models (Povey et al., 2018) using the Kaldi toolkit (Povey et al., 2011). The architecture for the TDNN-F models corresponds to that in the librispeech egs[7] in Kaldi toolkit. We use 40-dimensional MFCCs together with 100-dimensional i-vectors as feature parameters. All the manually transcribed acoustic data was speed perturbed 3-ways (speeds of 0.9, 1.0, and 1.1) (Ko et al., 2015) before acoustic model training. The word error rate for different test sets using TDNN-F acoustic models trained separately for each language are shown in Table 4. The acoustic models are trained with lattice-free MMI followed by discriminative training. For Inuktut, we also trained a bidirectional long short-term memory (BLSTM) acoustic model (Graves et al., 2013) from Inuktut training data as we got the best results for Inuktut with BLSTM acoustic models (compare lines 4 and 5 in Table 4).

The WER with unilingual acoustic models in Table 4 is compared with WER for multilingual models in the next section.

Language	Dev set	WER
1. Kurmanji Kurdish	OpenASR20	65.4%
2. Kurmanji Kurdish	TWB-audio	58.5%
3. Cree	CBC Cree 6files	62.5%
4. Inuktut	Inuktut 7files	87.9%
5. Inuktut (BLSTM)	Inuktut 7files	78.2%

Table 4: *WER for development sets for Kurmanji Kurdish, Cree and Inuktut with separate TDNN-F acoustic model trained for each language. For Inuktut, line 5 gives results with BLSTM acoustic models also.*

3.2. Multilingual Acoustic Model Training with 12 Languages

We trained multilingual DNNs from the acoustic training data from all the 12 languages as described in the *data resource* section. The total audio is roughly 429 hours (339 hours manually transcribed and 90 hours automatically transcribed) and the manually transcribed audio is speed perturbed 3 ways (Ko et al., 2015) before training. To optimise acoustic modeling, we tried two different phone sets: one phone set where the non-silence phones for each language are kept separate by adding a language tag to each phone (a total of 522 phones), and second phone set where all the common phones from all the languages are merged (a total of 193 phones). For each of these phone sets, we compared two different acoustic models (TDNN-F and BLSTM). The WER for the different development sets with multilingual training with separate phones per language is shown in Table 5, while the WER with merged common phones from all the 12 languages is shown in Table 6.

The best results from all four scenarios (separate phones for each language versus merged phones, TDNN-F versus BLSTM) are shown in bold in Tables 5 and 6. We can see that it is not the same acoustic model training scenario that gives the best results for all four different development sets. We see that separate phones with TDNN-F acoustic models give the lowest WER for Kurmanji Kurdish (develop-

[7] https://github.com/kaldi-asr/kaldi/egs/librispeech/s5/local/chain/run_tdnn.sh

ment sets OpenASR20 and TWB-audio). For Cree dev set, the lowest WER is with TDNN-F acoustic models with merged phones (62.1% with merged phones versus 62.8% with separate phones). For Inuktut dev set, the lowest WER is with the BLSTM acoustic models with separate phones, and the WER is significantly lower compared to TDNN-F acoustic models (71.1% WER for BLSTM versus 78.1% for TDNN-F).

Dev set	TDNN-F	BLSTM
OpenASR20	**64.9%**	75.7%
TWB-audio	**50.2%**	74.9%
CBC Cree 6files	62.8%	70.1%
Inuktut 7files	78.1%	**71.1%**

Table 5: *WER for development sets for Kurmanji Kurdish, Cree and Inuktut with multilingual acoustic models trained with separate phones for each language. We compare TDNN-F acoustic model versus BLSTM acoustic models in this scenario. WER in bold is the best WER from all training scenarios.*

Dev set	TDNN-F	BLSTM
OpenASR20	65.5%	76.1%
TWB-audio	66.0%	79.3%
CBC Cree 6files	**62.1%**	73.2%
Inuktut 7files	89.6%	79.8%

Table 6: *WER for development sets for Kurmanji Kurdish, Cree and Inuktut with multilingual acoustic models trained with merged phones from all 12 languages. We compare TDNN-F acoustic model versus BLSTM acoustic models in this scenario also. WER in bold shows the best WER from all training scenarios.*

To see why merged phones give lower WER for Cree and higher WER for Inuktut and Kurmanji Kurdish, we computed number of phones that do not overlap other languages (column 2, Table 7), and the average number of overlapping languages for other phones (column 3). From Table 7 we see that Cree has 4 phones that do not overlap with phones from other languages. The other phones in Cree overlap with 7.4 other languages on an average. So these statistics are not different enough to explain the differences in WER. The only thing that stands out is the fact that Cree has larger training audio (101 hours) (Table 1) than Inuktut (78 hours) and Kurmanji Kurdish (88 hours), so acoustic models with merged phones are dominated by Cree, reducing WER for Cree while increasing WER for other languages.

If we compare these lowest multilingual word error rates (WER) with WER from unilingual training of TDNN-F acoustic models for each language, then we see that multilingual training reduces WER for each dev set. For OpenASR20 dev set, the WER goes down from 65.4% to 64.9% (0.5% reduction in WER absolute), for TWB-audio, the WER goes down from 58.5% to 50.2% (8.3% reduction in WER absolute). For *CBC Cree 6files* dev set, the WER goes down from 62.5% to 62.1%, and for *Inuktut 7files* dev set the WER goes down from 78.2% to 71.1%.

Language	Phones unique to language	Overlap with # of languages
Cree	3_r 5 T V	7.4
Inuktut	K N: R	8.6
Kurmanji		6.7

Table 7: *Phones unique to the language (column 2) and average number of languages with which the remaining phones are shared (column 3).*

3.3. Incremental Training per Language from Multilingual Acoustic Models

We found that by starting with the final multilingual acoustic models as the initial models, and training them with one more epoch from just the acoustic data from one language with a small learning rate, we can reduce the WER even further. This reduction in WER is discussed in this section.

We tried many variations of incremental training of TDNN-F acoustic models with just one language data starting with the already trained multilingual TDNN-F acoustic models: we varied the learning rate, the number of epochs to train, and number of previous models to combine to produce the final model. We found that if we just train for 1 more epoch with just the training data from the language and combine models so that the original multilingual model is included in the combined model, then we get the lowest WER. It seems to be important to include the multilingual model in the model combination. The incremental training is followed by discriminative training of the models with just acoustic data from that language. The WER with this incremental training for multilingual TDNN-F acoustic models with separate phones for each language is shown in Table 8. From this Table, we see that, for each development set, there is a significant reduction in WER, especially for TWB-audio (2.2% absolute) and for *CBC Cree 6files* development set (2.5% absolute).

Dev set	Before incremental training	After incremental training
OpenASR20	64.9%	64.4%
TWB-audio	50.2%	48.0%
CBC Cree 6files	62.8%	60.3%
Inuktut 7files	78.1%	76.9%

Table 8: *WER for development sets for Kurmanji Kurdish, Cree and Inuktut with incremental training per language of multilingual TDNN-F acoustic models with separate phones for each language. We compare TDNN-F acoustic models before and after incremental training.*

Since the TDNN-F multilingual acoustic models with merged phones gave the best results for *CBC Cree 6files* dev set (see Table 6), we trained these acoustic models incrementally with only Cree data. Similarly, since BLSTM multilingual acoustic models with separate phones gave the lowest WER for *Inuktut 7 files* dev set (see Table 5), we incrementally trained this model with only Inuktut training data. The results are shown in Table 9. From this Table we see that the WER is reduced for both the Cree and Inuktut development sets after incremental training of acoustic models.

From Tables 8 and 9 we see that incremental training resulted in reduced WER in every scenario we tried: for both TDNN-F models with separate phones or merged phones, and for BLSTM models with separate phones.

Dev set	Before incremental training	After incremental training
CBC Cree 6files	62.1%	60.7%
Inuktut 7files	71.1%	69.6%

Table 9: *WER for Cree development set after incremental training of multilingual TDNN-F model with merged phones from Cree data alone, and for Inuktut development set after incremental training of multilingual BLSTM acoustic models with separate phones from Inuktut data alone.*

From Tables 8 and 9 we see that for Inuktut, the WER with TDNN-F models (76.9%) is much higher than the WER with BLSTM acoustic models (69.6%). Inuktut is the only language where this is true. For other languages the reverse is true: the WER with TDNN-F acoustic models is lower than that for BLSTM acoustic models. Also, Inuktut is the only language where we use syllable-based decoding instead of word-based decoding. To see that if syllable-based decoding has something to do with this issue, we ran word based decoding for Inuktut with a dictionary of 129k words (Gupta and Boulianne, 2020a). The decoding results with this dictionary are shown in Table 10. We see that with word-based dictionary TDNN-F acoustic models give lower WER (106.6%) compared to BLSTM acoustic models (121.6%). The reason for the high WER is because the 129k dictionary still leads to 62% out-of-vocabulary words in the Inuktut dev set leading to very high substitution rates (see Table 10 for breakdown).

To understand why syllable-based decoding for Inuktut results in higher word error rate with TDNN-F models, we compare decoding results for TDNN-F models versus BLSTM acoustic models in Table 11. From this Table we see that there are significantly more word deletions (19.6%) for TDNN-F acoustic models compared to BLSTM based acoustic models (9.2%). To understand these word deletions, we need to understand how word boundaries are detected from decoded syllable sequences in syllable-based decoding.

Acoustic model	TDNN-F	BLSTM
insertion	6.6%	21.6%
deletion	8.7%	1.1%
substitution	91.3%	98.9%
WER	106.6%	121.6%
total decoded words	7520	9251

Table 10: *WER breakdown into percent insertion, deletion and substitution for Inuktut dev set with word-based decoding. We also show the total number of decoded words in the decoded ctm file.*

The dictionary contains a total of 3158 syllables. The syllables can occur in beginning, middle or end of the word. So syllables at the start of the word are marked with B_, while syllables at the end of the word are marked with _E markers. The total number of syllables includes these marked syllables. The language model is trained with text in syllables that include these marked syllables. The syllables themselves, whether marked or not, have the same phonetic transcription. So word boundary markers are decoded using the language model weights, and any acoustic cues for word boundaries in the audio. In TDNN-F based decoding, only every third frame is used for decoding, so some of these acoustic cues may be missing more than in BLSTM based decoding which uses every frame for decoding. That is why we see 9.2% missing word boundaries for BLSTM based decoding versus 19.6% missing word boundaries for TDNN-F based decoding (see deletion row in Table 11).

Acoustic model	TDNN-F	BLSTM
insertion	1.3%	3.0%
deletion	19.6%	9.2%
substitution	56.0%	57.5%
WER	76.9%	69.6%
total decoded words	6270	7205

Table 11: *WER breakdown into percent insertion, deletion and substitution for Inuktut dev set with syllable-based decoding. We also show the total number of decoded words in the decoded ctm file.*

In Table 12 we compare WER for best unilingual training versus the best multilingual training (multilingual training followed by incremental training). We see that even with over 40 hours of acoustic training data for each language, multilingual training gives significant reduction in WER for each language. For *TWB-audio*, the WER goes down by 10.5% (absolute), while for *Inuktut 7 files*, the WER goes down by 8.6% (absolute).

3.4. Pretrained XLSR-based Speech Feature Parameters

We extracted speech representations for all our multilingual acoustic data (training and development) using the publicly available XLSR-53 model (Conneau et al., 2020), a wav2vec2.0 model pre-trained on 56k hours of speech in 53 languages, from 3 datasets:

Dev set	Unilingual training	Multilingual training
OpenASR20 (Kurmanji)	65.4%	64.4%
TWB-audio (Kurmanji)	58.5%	48.0%
CBC Cree 6files	62.5%	60.3%
Inuktut 7files	78.2%	69.6%

Table 12: *WER for the best scenario for unilingual versus multilingual training for the different development sets.*

MLS (Multilingual LibriSpeech), CommonVoice and Babel[8]. XLSR is pretrained directly from raw unlabeled speech audio. We used the encoder output representations z_t which are 512-dimensional vectors computed every 20 ms. We added these features to our 40-dimensional MFCC features for training the TDNN-F acoustic models. Since the frame rate for the 40-dimensional MFCC features was 10 ms, we duplicated every frame of the 512-dimensional features in order to get the same frame rate before concatenating the two features. The WER with these joint features for multilingual TDNN-F models with separate phones for each language is given in Table 13, and the WER with/without these XLSR-based features is compared in Table 13. The joint features gave lower WER for 2 out of the 4 dev sets (shown in bold). We probably need to adapt the pre-trained XLSR model to our multilingual training set in order to get significant reduction in WER.

Dev set	40-dim MFCC	40-dim MFCC + 512-dim wav2vec2.0
OpenASR20	64.9%	**64.3%**
TWB-audio	50.2%	51.8%
CBC Cree 6files	62.8%	**62.1%**
Inuktut 7files	78.1%	79.0%

Table 13: *WER for development sets for Kurmanji Kurdish, Cree and Inuktut with multilingual acoustic models trained with separate phones for each language. We compare TDNN-F acoustic model with 40-dim MFCC versus 40-dim MFCC + 512-dim pretrained XLSR-based features.*

4. Discussion of Results

When we see the word error rates (WER) for the different development sets, we see a big variation in WER, from 48% for TWB-audio to 69.6% for Inuktut dev set. We would like to explain some of these differences. Let us take Kurmanji first: There is a 16.3% difference in WER between OpenASR20 dev set (64.3%) and TWB-audio (48%). There are multiple reasons for this difference in WER. First, the OpenASR20 dev set corresponds to conversational speech with many hesitations, repeats, noise, pauses, etc. So a voice activity detector (VAD) separates noise segments from speech segments and we recognize only the speech segments. This VAD helps reduce WER, but still the WER is higher than that for read speech. Secondly, the topics of discussion can vary significantly and it is difficult to train a good language model for conversational speech from any kind of news corpus. The only text corpus that reduced WER for the OpenASR20 dev set was the transcribed conversations in the training set. There could be two reasons for it. One is probably that the conversation topics may be somewhat similar to those in the development set. Secondly, the training and dev sets were transcribed by the same group of transcribers with the same instructions, so the transcribed text is probably quite uniform both in terms of spelling and putting accent marks on words. When we compare the percentage of words accented in the LDC training text for Kurmanji versus the percentage of accented words in TWB-news and TWB-other, we see that 40.9% words are accented in LDC text versus 45.7% in TWB-news, and 46.9% in TWB-other. We also find that when we add TWB-news or TWB-other text to the language modeling text, the WER for OpenASR20 dev set goes up significantly. All the above issues lead to high WER for OpenASR20 dev set.

The TWB-audio in Kurmanji Kurdish is different: the text corresponds to Kurmanji Kurdish tales, and they have been read by one speaker. There are no hesitations or repeats, and only small silent gaps between words. The WER with/without voice activity detector is about the same. Also, the language model created from over 45 million words of text from TWB-news and TWB-other reduces the WER significantly for the TWB-audio. So the language model created from TWB-news and TWB-other more closely represents the word contexts in TWB-audio. For all the above reasons, we get the lowest WER for the TWB-audio when compared to the other dev sets.

The Cree dev set is extracted from CBC (Canadian Broadcasting Corporation) broadcasts in Cree from four different current affair programs. The recordings are clean except that for musical intervals and telephone interviews. The training set for Cree is also from CBC broadcasts. So the language model created from these broadcasts does represent the dev set. For that reason, the WER for Cree is reasonable: around 60%. The WER is higher than the 48% WER for TWB-audio for a few reasons: music and telephone interviews increases the WER, the broadcasts are not read speech, and the language model for Cree is trained from only 245k words of text, while the language model for TWB-audio is trained from over 45 million words of text.

The Inuktut dev set is extracted from recorded stories and sometimes include singing and chanting. Like Cree, both the training and dev sets are from the same source. For Inuktut, we also have a million words of

[8] These datasets do not include Cree or Inuktut, but Babel includes Kurmanji Kurdish.

text from Nunavut parliament proceedings. However, the context for the parliament proceedings is quite different from the stories recited by elders in the training and dev sets. But the major reason why Inuktut dev set WER of 69.6% is much higher than the 60.3% for Cree is that Inuktut is highly polysynthetic, more so than Cree. Even with a word dictionary of a million words, the out-of-vocabulary rate for unseen text can be as high as 60% (Gupta and Boulianne, 2020a). For that reason, we use a syllabic dictionary and decode the audio as a sequence of syllables. The syllables have word boundary markers. Those markers in the decoded syllables are used to convert decoded syllable sequence into word sequence. The use of syllables for decoding and then using word markers to convert syllable sequences into word sequences increases the word error rate.

5. Conclusions

We experimented with multilingual speech recognition using acoustic data from 12 low resource languages (10 languages from OpenASR20 challenge, plus Canadian Indigenous languages Cree and Inuktut) in order to optimize recognition accuracy for three low resource languages: Kurmanji Kurdish, Cree and Inuktut. We tried two different multilingual phone sets: separate phones for each language, and a phone set where common phones are merged. We find that for Kurmanji Kurdish and for Inuktut, keeping separate phone sets is better, while for Cree merged phone set gives lower word error rate (WER). The reason for this is that Cree has the largest acoustic training data, and merged phone set acoustic models represent Cree more than the other languages.

We also compared WER with factored time delay neural networks (TDNN-F) and bidirectional long short-term memory (BLSTM) neural networks. The TDNN-F based acoustic models give lower word error rates for both Cree and Kurmanji Kurdish, while BLSTM acoustic models give significantly lower WER for Inuktut. It also happens that Inuktut is decoded with syllables to provide sufficient coverage of unseen text, and decoded syllable sequences are transformed into words by syllables with word-end markers. We show that syllable-based decoding is the reason why BLSTM acoustic models give lower WER than TDNN-F based acoustic models.

We also show that training the multilingual acoustic models for one more epoch with just the acoustic data from one language results in significant reduction in WER. In this scenario, the final acoustic model, which is a combination of previous N models, should include the final model trained with multilingual data.

We concatenated 512-dimensional features from pre-trained XLSR models with 40-dimensional MFCC features, and the combined features gave a small reduction in WER for two development sets.

To summarize, we get significant reduction in WER with multilingual acoustic models as compared to unilingual acoustic models. We show when we can use merged phone sets and when we should use separate phone sets. We also show why we should use BLSTM acoustic models for highly polysynthetic languages where we need to decode using syllabic or sub-word dictionary. Training for one more epoch (with multilingual models as initial models) with acoustic data for only one language leads to significant reduction in word error rate for that language.

We also discuss why the word error rate (WER) for different development sets varied from 48% (for TWB-audio) to 69.6% (for Inuktut). The WER depends on acoustics (read speech is the easiest, followed by broadcast stories, and then conversational speech). It also depends on how well the context of the development audio matches the context of language modeling text. The conversational speech text in general is quite different from the news text generally available over the internet. Also, highly polysynthetic languages like Inuktut are difficult to decode, since they require a sub-word dictionary and a way to merge decoded sub-word sequences into word sequences. Ultimately, this work stresses the importance of including morphologically complex languages in multilingual research if we expect conclusions to generalize to all languages.

6. Acknowledgements

The authors would like to thank Ministry of Economy and Innovation (MEI) of the Government of Quebec for the continued support.

We would also like to thank Alp Öktem from Translators without Borders (TWB) for providing us with transcribed acoustic data (TWB-audio) for a second development set (in Kurmanji Kurdish) and over 46 million words of text in Kurmanji Kurdish from various sources for language modeling.

7. Bibliographical References

Bentz, C., Ruzsics, T., Koplenig, A., Samardži, T., and Samardži ́c, T. S. (2016). A Comparison Between Morphological Complexity Measures: Typological Data vs. Language Corpora. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 142–153.

Chen, G., Khudanpur, S., Povey, D., Trmal, J., Yarowsky, D., and Yilmaz, O. (2013). Quantifying the value of pronunciation lexicons for keyword search in low resource languages. In *Proc. ICASSP*, pages 8560–8564.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2020). Unsupervised Cross-lingual Representation Learning for Speech Recognition. *arXiv eprint*. arXiv:2006.13979.

Dalmia, S., Sanabria, R., Metze, F., and Black, A. W. (2018). Sequence-based Multi-lingual Low Resource Speech Recognition. *arXiv eprint*. arXiv:1802.07420.

Gales, M. J. F., Knill, K. M. ., Ragni, A., and Rath, S. P. . (2014). Speech Recognition and Keyword Spotting for Low Resource Languages: BABEL project research at CUED. In *Proc. SLTU*, pages 14–16.

Graves, A., Jaitly, N., and Mohamed, A.-R. (2013). Hybrid speech recognition with Deep Bidirectional LSTM. In *Proc. ASRU*, pages 273–278, Olomouc, Czech Republic.

Gupta, V. and Boulianne, G. (2020a). Automatic Transcription Challenges for Inuktitut, a Low-Resource Polysynthetic Language. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 2521–2527.

Gupta, V. and Boulianne, G. (2020b). Speech transcription challenges for resource constrained indigenous language Cree. In *Proceedings of the 1st Joint SLTU and CCURL Workshop*, pages 362–367.

Huang, J.-T., Li, J., Yu, D., Deng, L., and Gong, Y. (2013). Cross-Language Knowledge Transfer Using Multilingual Deep Neural Network with shared hidden layers. In *Proc. ICASSP*, pages 7304–7308.

Kannan, A., Datta, A., Sainath, T. N., Weinstein, E., Ramabhadran, B., Wu, Y., Bapna, A., Chen, Z., and Lee, S. (2019). Large-Scale Multilingual Speech Recognition with a Streaming End-to-End Model. *arXiv eprint*. arXiv:1909.05330.

Knill, K. M., Gales, M. J., Ragni, A., and Rath, S. P. (2014). Language independent and unsupervised acoustic models for speech recognition and keyword spotting. In *Proc. Interspeech*, pages 16–20.

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Proc. Interspeech*, pages 3586–3589.

Kuhn, R., Davis, F., Désilets, A., Joanis, E., Kazantseva, A., Knowles, R., Littell, P., Lothian, D., Pine, A., Running Wolf, C., Santos, E., Stewart, D., Boulianne, G., Gupta, V., Maracle Owennatékha, B., Martin, A., Cox, C., Junker, M.-O., Sammons, O., Torkornoo, D., Thanyehténhas Brinklow, N., Child, S., Farley, B., Huggins-Daines, D., Rosenblum, D., and Souter, H. (2020). The indigenous languages technology project at NRC Canada: An empowerment-oriented approach to developing language software. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5866–5878.

Li, X., Dalmia, S., Li, J., Lee, M., Littel, P., Yao, J., Anastasopoulos, A., Mortensen, D. R., Neubig, G., Black, A., and Metze, F. (2020). Universal phone recognition with a multilingual allophone system. *arXiv eprint*. arXiv:2002.11800v1 [cs.CL].

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlícek, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., and Veselý, K. (2011). The Kaldi Speech Recognition Toolkit. In *Proc. ASRU*.

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohamadi, M., and Khudanpur, S. (2018). Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Proc. Interspeech*, pages 3743–3747.

Schwartz, L., Tyers, F., Levin, L., Kirov, C., Littell, P., Lo, C.-k., and Prud'hommeaux, E. (2020). Final Report of the Frederick Jelinek Memorial Summer Workshop on Neural Polysynthetic Language Modelling. *arXiv eprint*. arXiv:2005.05477.

Trmal, J., Chen, G., Povey, D., Khudanpur, S., Ghahremani, P., Zhang, X., Manohar, V., Liu, C., Jansen, A., Klakow, D., Yarowsky, D., and Metze, F. (2014). A keyword search system using open source software. In *Proc. SLT Workshop*, pages 530–535.

Vu, N., Povey, D., Motlicek, P., Schultz, T., and Bourlard, H. (2014). Multilingual deep neural network based acoustic modeling for rapid language adaptation. In *Proc. ICASSP*, pages 7689–7693.

Zhang, X., Trmal, J., Povey, D., and Khudanpur, S. (2014). Improving deep neural network acoustic models using generalized maxout networks. In *Proc. ICASSP*, pages 215–219.