Multilingual models with language embeddings for low-resource speech recognition

Léa-Marie Lam-Yee-Mui[1,2], Waad Ben Kheder[2], Viet-Bac Le[2], Claude Barras[2], Jean-Luc Gauvain[1]

[1]University of Paris-Saclay, CNRS, LISN, France
[2]Vocapia Research, France
lea-marie.lam-yee-mui@lisn.upsaclay.fr

Abstract

Speech recognition for low-resource languages remains challenging and can be addressed with techniques such as multilingual modeling and transfer learning. In this work, we explore several solutions to the multilingual training problem: training monolingual models with multilingual features, adapting a multilingual model with transfer learning and using language embeddings as additional features. To develop practical solutions we focus our work on medium size hybrid ASR models. The multilingual models are trained on 270 hours of iARPA Babel data from 25 languages, and results are reported on 4 Babel languages for the Limited Language Pack (LLP) condition. The results show that adapting a multilingual acoustic model with language embeddings is an effective solution, outperforming the baseline monolingual models, and providing comparable results to models based on state-of-the-art XLSR-53 features but with the advantage of needing 15 times fewer parameters.

Index Terms: speech recognition, low-resource languages, multilingual modeling, language embedding

1. Introduction

In recent years, multilingual approaches have emerged as a promising solution for improving speech recognition accuracy in low-resource languages, leveraging the shared linguistic features across multiple languages to overcome data scarcity and resource limitations. Multilingual acoustic models training has previously been explored by pooling data coming from different languages and using different criteria such as Lattice-Free Mutual Maximum Information (LF-MMI) [1], End-to-End cross-entropy-based models [2, 3] and meta learning [4] showing the potential of multilingual pre-training. While these models generally outperform monolingual models trained on limited amounts of data, they can still be improved by using some task-specific data. This problem is usually performed using transfer learning [5] with a monolingual labeled data set as a target. Besides being used as adaptable acoustic models, multilingual networks can also be used as a feature extractor where bottleneck features are computed and fed as input to help make hybrid HMM/DNN models more robust. The XLSR-53 model [6] is an example of such models based on the large version of Wav2vec2.0 architecture [7] and trained with a constrastive loss, learning to recognize latent speech representations. XLSR-53 is trained on about 50k hours of untranscribed training data in 53 languages and comprises 300 million parameters. In the context of speech recognition in low-resource settings, it is common to add layers on top of the last layer. The weights of the added layers are estimated with a supervised loss function, such as LF-MMI [8] or Connectionist Temporal Classification (CTC) [6]. Moreover, speech representations

from XLSR-53 can also be used as multilingual features to train a monolingual model. These approaches are state-of-the-art for speech recognition for low-resource languages. However, XLSR-53 models are very large and complex, requiring significant computational resources to train and deploy. This can make it difficult or expensive to use these models in real-world applications.

An alternative solution is to provide a language embedding as input in order to help condition the ASR system's output on language-specific features. This can be achieved by providing a one-hot encoding vector as "conditional input" in order to specify the language of each speech segment [9]. A second solution is to use an acoustic language embedding (such as x-vector or d-vector) which can be viewed as "segmental features" that carry information related to the acoustic characteristics of each language. Language embeddings have been previously integrated with multilingual DNN acoustic models, trained with cross-entropy [10, 11] and LF-MMI [12] and End-to-End models [13]. [12] uses in particular x-vector-based language embeddings in an unified system for speech recognition and language identification. Besides using speaker [14] and language as targets, accent embeddings have also been explored in [15, 16] to improve the acoustic models performance in the context of multi-accent ASR and showed that segment-level embeddings is able to capture high-level accent-related information.

In this paper, we compare solutions to the multilingual modeling problem, with different types of input features and the use of x-vector embeddings to capture language-related information. We also evaluate language x-vector against language one-hot encoding vector. We evaluate these solutions in tandem with a large Wav2Vec2.0 (XLSR-53) and a compact TDNN-F multilingual models using the iARPA Babel data set in order to show the potential of each component in the context of very limited training data.

2. Data sets

In this section, we present the data sets used to train and evaluate the developed ASR models.

2.1. Training data

We use the 8kHz Conversational Telephone Speech data from the iARPA Babel program [17], which is spontaneous speech collected for 25 low-resource languages in several acoustic environmental conditions. We focus on two different low-resource conditions:

- The first setting uses the Limited Language Packages (LLP) of Babel to study the effect of the developed techniques on low-resource languages. The LLP contain around 10 hours

for each language and we will refer to this amount of hours to signify LLP conditions for the remaining of the paper.

- A second setting corresponding to a very low-resource condition obtained by randomly selecting a subset of 2.5 hours from the LLP training data for each of the target language.

For each condition, we limit the amount of text data to train the language models to only the transcriptions of the audio data (i.e. 10h or 2.5h). This means that the only available data for each language is the limited acoustic data and its manual transcripts. Such conditions are quite common when developing ASR systems for very low resource languages.

2.2. Test data

All experiments were evaluated using the official Babel development sets for Amharic (amh), Assamese (asm), Georgian (geo) and Kurmanji (kmr). The target languages were chosen so that they cover several language families (Afro-Asiatic, Indo-European and Kartvelian) and characteristics to validate our approach in a variety of languages.

2.3. Data augmentation

The original Babel data are augmented with 3-fold speed perturbation, data reverberation and addition of background noises (point-source noises, real room impulse responses and isotropic noises [1]).

3. Methods

3.1. Multilingual models

Building a multilingual acoustic model can help overcome the constraints due to the limited amount of training data available for low-resource languages by leveraging common phonemic features across different languages.

The multilingual models in this work are trained with two multilingual data sets. The first one uses all 25 of the Babel languages under the LLP conditions, for a total of 270 hours. The second multilingual model uses all Babel languages (LLP condition) after excluding the four test languages, for a total of 220 hours.

Training such multilingual models requires the careful selection of a shared phone set which should, not only, take full advantage of the available Babel data and accurately cover any language's segmental inventory, but also generalise well to other "unseen" languages (universal phone set).

3.1.1. Phone set definition

The PHOIBLE [18] repository contains the phonological inventories for over 2,000 languages (including the ones present in Babel) and is a valuable resource while defining the phone set by paying particular attention to the phones/allophones frequencies in each language. The full procedure to define the multilingual phone set is as follows:

1. **Building a common phonetic inventory:** We use Epitran [19] and Espeak-ng [2] to generate phonetic lexicons based on the IPA for all supported languages (multiple pronunciations are kept for languages present in both Epitran and Espeak-ng). Due to the high level of phonetic granularity provided by these tools, the resulting phone set on 25 languages can be quite large (more than 100 phones) and requires reduction.

2. **Reducing the phone set:** The primary phone set can be reduced in different ways. First, suprasegmental qualities such as stress, length, pitch and tone are stripped away. Then, implosive and ejective variants of each consonant are combined. These steps are carefully carried out after checking the most frequent phones for each language on PHOIBLE and avoiding the combination of frequent discriminative phones.

3. **Filtering rare phones:** Based on the PHOIBLE phonological inventory of each Babel language, the rare phones (where the representation is lower than 5%) are mapped to the closest (and more frequent) phones remaining from the previous step.

For the four languages unsupported by Epitran/Espeak-ng (Dholuo, Haitian Creole, Igbo and Mongolian) the phonetic dictionaries provided by iARPA as part of the Babel project are used and their phones are mapped to the final multilingual phone set. The resulting phone set used in the remaining of the paper contains 70 phones in total; 67 speech phones and 3 non-speech phones representing silence, filler words and breath noise.

3.1.2. Model Architecture and Training

The multilingual acoustic models are based on the Factorized Time Delay Neural Network (architecture TDNN-F) [20] and trained with the augmented data for 4 epochs . They are composed of 15 TDNN-F layers. The baseline models use 41-dimensional PLP features which are spliced (x3) and transformed with an LDA.

The alignments for the final TDNN-F model training (with augmented data) come from another TDNN-F model trained for 6 epochs with alignments of non-augmented data obtained with a standard HMM/GMM model. Kaldi toolkit [21] is used for the neural network training with the LF-MMI loss function [22].

3.2. Language embedding

To help the training of the multilingual acoustic model, linguistic information can be added by concatenating a language-related vector to the input features for each frame.

3.2.1. Language one-hot encoding

The language identity can be given through a fixed-length vector. However, it requires all target languages to be seen during the training. In particular, a 25-dimensional one-hot encoder vector is created to indicate which language the utterance is spoken in. During inference on a target language, the one-hot encoder corresponding to the language to decode is provided to the model.

3.2.2. Language x-vector

For speech recognition for an unseen language, the one-hot encoder is no longer usable. A learned language embedding can replace the one-hot encoder vector in the inputs of the multilingual acoustic model and hopefully carry enough implicit linguistic information about the unseen language.

For this work, we choose to use the x-vector as language embedding. Mostly used for speaker recognition [23], the x-vector model consists of a feed-forward DNN/TDNN that maps sequences of variable-length speech features to fixed-dimensional embeddings. In the context of language adaptation, the x-vector extractor is trained to discriminate the languages. The architecture of the x-vector extractor used in our work is

---

[1] https://www.openslr.org/28/
[2] https://github.com/espeak-ng/espeak-ng

similar to the one used for speaker identification as described in [24]. The language x-vector extractor has 3.2 million parameters. It is trained for 5 epochs with a learning rate ranging from 0.001 to 0.0001. The training data set is identical to the training data set for the multilingual acoustic models but is automatically segmented to remove audio segments without speech. Data augmentation is also applied.

To improve the robustness of the language x-vector, the pooling layer computes the statistics (mean) for all speech turns of the same speaker. A 128-dimensional language embedding is thus obtained for each speaker.

3.3. Transfer learning

To improve the accuracy of the multilingual model for a given target language, a transfer learning step is added using limited language adaptation data. Transfer learning can be done by continuing the training of the multilingual model with the target language data. Alternatively, as proposed in [5], one can replace the last few layers of the multilingual model by randomly initialized layers with the last one corresponding to the output of a monolingual model. We refer to this last method as adaptation for the remaining of the paper. For this work, one to three final layers of the multilingual acoustic model are replaced and trained from a random initialization.

Different parameter settings can be explored while performing transfer learning such as the learning rate and the number of epochs which can vary depending on the amount of target data available. Other parameters such as l_2 regularization constant can also be used to control the penalty imposed on the loss function and prevent the model from overfitting to the source task and help the model generalize better to the target task. The learning rate factor is another parameter that should be chosen carefully to control the rate at which the parameters of the neural network model are updated during training. During adaptation, we experimented with a range of learning rates (from 5.10^{-5} to 1.10^{-3}) multiplied by different factors according to the updated layers (0.1 or 1.0), to avoid losing too much information from the multilingual pre-training. We adapted the model for 3 epochs.

3.4. Features

The multilingual phonetic TDNN-F acoustic model can also be used as a feature extractor. The extracted bottleneck features are then fed to a separate acoustic model. This allows the ASR model to benefit from the high-level, compact representations of speech signals provided by the bottleneck features, while also allowing the model to be trained on a smaller data set, therefore requiring less computational resources. The bottleneck features are extracted from the 15th layer of the multilingual model, and used to train a new monolingual TDNN-F model, using the alignments obtained by the original monolingual model trained on the non-augmented data.

We compare these bottleneck features to features extracted from self-supervised models. Here we only consider features extracted from XLSR-53 to keep the number of parameters for the full pipeline of the ASR system reasonable and the computational resources at a minimum. For the sake of keeping the used resources as low as possible, features from the XLS-R models [25] have not been used even if we expect an improvement with these features, thanks to the additional data and number of parameters present in these bigger versions of XLSR-53.

The S3PRL toolkit [26] is used to extract the 1024-dimensional features from the last layer of the XLSR-53 model

and PyKaldi [27] is used to convert the features to Kaldi format.

4. Results

The experimental results are presented in Table 1 for the 4 targeted languages. The monolingual baselines WERs are given in rows (a) and (b) and correspond to models trained either with nominally 10h or with 2.5h of data from the iARPA Babel corpus. The TDNN-F acoustic models include respectively 11 million and 8 million parameters, with the model sizes chosen to minimize the WERs. For both conditions we use a 3-gram language model trained only on the manual transcripts of the audio training data. The language model data is limited to the available transcriptions for each condition, even though work in the Babel program showed that using additional external texts generally improved recognition performance, in particular by reducing the out-of-vocabulary rate. The average WERs over the 4 languages are 59.6% and 84.3% for the 10h and 2.5h training conditions, respectively.

It should be be noted that better results have been reported on the same data in the Babel program [28] and more recently for the OpenASR21 [3] challenge [29, 30]. However these results were obtained using considerably more language model training data than just the acoustic data transcripts. Using more texts for the LM training we easily lower the WER to under 40% on Amharic with the same acoustic models.

The results with the multilingual acoustic models are given in rows (c) to (l), where here the training data used to build the baseline models is used for adaptation. To allow for meaningful comparisons, the size of the acoustic model is given for each condition. The multilingual acoustic models are trained as described in section 3.1, using either a set of 25 languages including the 4 target languages, or on the remaining 21 languages after excluding the 4 target languages. In all cases, the language model is trained on the transcripts of the adaptation data of the target language (10h or 2.5h). The average WER using the unadapted multilingual acoustic models, without and with language embeddings (rows c and d), are respectively 57.8% and 55.0%. It can be seen that multillingual model outperforms the monolingual one by 4.6% absolute and also demonstrates the effectiveness of the language embeddings (a 2.2% absolute gain).

Adapting the multilingual acoustic models with the 10h target language data reduces the average WER from 55.0% (row d) to 52.1% (row f). Two adaptation conditions were considered: adaptation with and without language embeddings, both with changing the last layers of the network. Adaptation results are also given using only 2.5h of target-language data, resulting in an average WER of 59.4% (row h). However, the experimental results with 2.5h of adaptation data for Amharic degrade and lower the average results, for a reason we have not yet understood.

Results comparing the language one-hot encoder and the language embedding for the multilingual models trained on 25 languages are given in Table 2. If the identity of the language is provided with the language one-hot encoder, the average WER of the unadapted multilingual model (57.8%) is reduced by an absolute 2.2% (row r) to 55.6%. A larger improvement is visible when adapting the multilingual model using the 10h adaptation data set (row u). The results also show that language x-vector outperforms the language one-hot encoder, without or with adaptation (rows (s) and (v) in the table).

---

[3] https://sat.nist.gov/openasr21

Table 1: *WER (%) of monolingual models, multilingual models without and with adaptation using 10h or 2.5h of target language data, and monolingual models using multilingual features. Results are given for 4 iARPA Babel languages: amharic (amh), assamese (asm), georgian (geo) and kurmanji (kmr) as well as an overall average.*

		# parameters	amh	asm	geo	kmr	average
	Monolingual acoustic and language models						
(a)	monolingual model trained with 10h	11.3 M	53.7	60,3	55.4	69.1	59.6
(b)	monolingual model trained with 2.5h	8.2 M	74.0	89.4	84.4	89.3	84.3
	Multilingual model (25 languages), LM trained on adaptation data (10h or 2.5h)						
(c)	no adaptation	18.5 M	54.0	58.3	54.6	64.3	57.8
(d)	with language embedding	18.7 + 3.2M	50.0	54.9	51.0	64.1	55.0
(e)	adaptation with 10h	18.3 M	48.5	52.7	48.5	63.0	53.2
(f)	with language embedding	18.3 + 3.2 M	47.3	51.4	47.6	62.0	52.1
(g)	adaptation with 2.5h	18.3 M	50.9	58.8	59.4	67.3	59.1
(h)	with language embedding	18.3 + 3.2 M	55.0	57.4	58.9	66.4	59.4
	Multilingual model (21 languages), LM trained on adaptation data (10h or 2.5h)						
(i)	adaptation with 10h	18.3 M	48.7	52.7	49.0	62.9	53.3
(j)	with language embedding	18.3 + 3.2 M	48.5	51.8	48.2	62.2	52.7
(k)	adaptation with 2.5h	18.3 M	52.0	59.2	60.5	68.0	59.9
(l)	with language embedding	18.3 + 3.2 M	56.1	58.8	59.7	67.1	60.4
	Monolingual models trained with multilingual features, LM trained on adaptation data (10h or 2.5h)						
(m)	features from PLP TDNN 10h	18.5 + 11.3 M	49.8	54.7	51.2	63.6	54.8
(n)	with language embedding	18.7 + 11.3 + 3.2 M	48.9	53.4	50.9	63.6	54.2
(o)	features from XLSR-53 10h	300 + 11.3 M	46.9	51.0	45.6	61.1	51.2
(p)	features from XLSR-53 2.5 h	300 + 8.2 M	61.1	72.3	63.9	61.2	64.6

Table 2: *Comparing one-hot language encoder vector to a language x-vector without and with adaptation. Average WER (%) for amh, asm, geo and kmr development sets of Babel, using multilingual models trained with 25 languages, including the target languages.*

(q)	multilingual model trained on 25 languages	57.8
(r)	with language one-hot encoder	55.6
(s)	with language embedding	55.0
(t)	adaptation with 10h	53.2
(u)	with language one-hot encoder	53.1
(v)	with language embedding	52.1

The results obtained with multilingual acoustic models trained on only 21 languages (excluding the 4 targeted languages) are given in rows (i) to (l) of Table 1. It can be seen that removing the target languages has only a small impact on the average WER (increasing it from 52.1% to 52.7%).

Results using multilingual features are given in rows (m) to (p) in Table 1. Models using the multilingual features (rows m and n) derived from the multilingual model perform less well than the adapted multilingual models (54.2% versus 52.1%). The use of XLSR-53 features for 10h training condition gives very competitive results: with an average WER of 51.2%, it is slightly better than the result obtained with the multilingual model (52.1%). However it is important to point out that the XLSR-53 is 15 times larger and that untranscribed training data includes the iARPA Babel FLP data for Georgian and Kurmanji.

The average WER for the 2 other languages are closer, 49.0% for XLSR-53 features (row o) versus 49.4% for the multilingual model (row h). For the 2.5h training condition, the multilingual model outperforms the XLSR-53 features (average 59.4% versus 64.6% across the 4 languages).

5. Conclusion

In this work we have compared acoustic models and training methods for low-resource speech recognition with the goal of keeping the models small and efficient. We found that the adaptation of a multilingual model with language embeddings appears to be a very effective solution to allowing the average WER on the 4 target languages to be reduced from 59.6% to 52.1% for the 10h training condition, and from 84.3% to 60.4% for the 2.5h training condition. The use of language embeddings increases the precision of the multilingual model both before and after adaptation. The use of multilingual XLSR-53 features yields comparable results, but at the cost of increasing the model size by a factor 15. It should also be noted that better results can be obtained and have been reported on these 4 languages by using all of the available Babel data for acoustic modeling with a semi-supervised training [28], adding other textual resources for language modeling and generating artificial textual data to further improve the LM [31]. ASR systems based on fine-tuning self-supervised models also perform better than our approach [6, 8] by benefiting from more textual data, but these approaches based on a better LM are not in the scope of this work which focuses on the acoustic modeling in low-resource speech recognition.

6. References

[1] S. R. Madikeri, B. K. Khonglah, S. Tong, P. Motlicek, H. Bourlard, and D. Povey, "Lattice-free maximum mutual information training of multilingual speech recognition systems." in *INTERSPEECH*, 2020, pp. 4746–4750.

[2] B. Li, R. Pang, T. N. Sainath, A. Gulati, Y. Zhang, J. Qin, P. Haghani, W. R. Huang, M. Ma, and J. Bai, "Scaling end-to-end models for large-scale multilingual asr," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 1011–1018.

[3] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein, and K. Rao, "Multilingual speech recognition with a single end-to-end model," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4904–4908.

[4] J.-Y. Hsu, Y.-J. Chen, and H.-y. Lee, "Meta learning for end-to-end low-resource speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7844–7848.

[5] P. Ghahremani, V. Manohar, H. Hadian, D. Povey, and S. Khudanpur, "Investigation of transfer learning for ASR using LF-MMI trained neural networks," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, Dec. 2017, pp. 279–286.

[6] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Interspeech 2021*. ISCA, Aug. 2021, pp. 2426–2430.

[7] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.

[8] M. Wiesner, D. Raj, and S. Khudanpur, "Injecting Text and Cross-Lingual Supervision in Few-Shot Learning from Self-Supervised Models," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 8597–8601.

[9] A. Kannan, A. Datta, T. N. Sainath, E. Weinstein, B. Ramabhadran, Y. Wu, A. Bapna, Z. Chen, and S. Lee, "Large-Scale Multilingual Speech Recognition with a Streaming End-to-End Model," in *Interspeech 2019*. ISCA, Sep. 2019, pp. 2130–2134.

[10] M. Müller, S. Stüker, and A. Waibel, "Language adaptive dnns for improved low resource speech recognition." in *Interspeech*, 2016, pp. 3878–3882.

[11] M. Müller, S. Stüker, and A. Waibel, "Language Adaptive Multilingual CTC Speech Recognition," in *Speech and Computer*, A. Karpov, R. Potapova, and I. Mporas, Eds. Cham: Springer International Publishing, 2017, vol. 10458, pp. 473–482, series Title: Lecture Notes in Computer Science.

[12] D. Liu, J. Xu, P. Zhang, and Y. Yan, "A unified system for multilingual speech recognition and language identification," *Speech Communication*, vol. 127, pp. 17–28, Mar. 2021.

[13] W. Hou, Y. Dong, B. Zhuang, L. Yang, J. Shi, and T. Shinozaki, "Large-Scale End-to-End Multilingual Speech Recognition and Language Identification with Multi-Task Learning," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 1037–1041.

[14] M. Karafiat, K. Vesely, J. H. Cernocky, J. Profant, J. Nytra, M. Hlavacek, and T. Pavlicek, "Analysis of X-Vectors for Low-Resource Speech Recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, ON, Canada: IEEE, Jun. 2021, pp. 6998–7002.

[15] A. Jain, M. Upreti, and P. Jyothi, "Improved Accented Speech Recognition Using Accent Embeddings and Multi-task Learning," in *Interspeech 2018*. ISCA, Sep. 2018, pp. 2454–2458.

[16] M. A. T. Turan, E. Vincent, and D. Jouvet, "Achieving multi-accent asr via unsupervised acoustic model adaptation," in *INTERSPEECH 2020*, 2020.

[17] M. Harper, *IARPA Babel Program*. [Online]. Available: http://www.iarpa.gov/index.php/research-programs/babel

[18] S. Moran and D. McCloy, Eds., *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History, 2019. [Online]. Available: https://phoible.org/

[19] D. R. Mortensen, S. Dalmia, and P. Littell, "Epitran: Precision g2p for many languages," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[20] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks," in *Interspeech 2018*. ISCA, Sep. 2018, pp. 3743–3747.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011.

[22] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Interspeech*, 2016.

[23] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[24] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors." in *Odyssey*, vol. 2018, 2018, pp. 105–111.

[25] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Interspeech 2022*. ISCA, Sep. 2022, pp. 2278–2282.

[26] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[27] D. Can, V. R. Martinez, P. Papadopoulos, and S. S. Narayanan, "Pykaldi: A python wrapper for kaldi," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018.

[28] V.-B. Le, L. Lamel, A. Messaoudi, W. Hartmann, J.-L. Gauvain, C. Woehrling, J. Despres, and A. Roy, "Developing STT and KWS systems using limited language resources," in *Interspeech 2014*. ISCA, Sep. 2014, pp. 2484–2488.

[29] G. Zhong, H. Song, R. Wang, L. Sun, D. Liu, J. Pan, X. Fang, J. Du, J. Zhang, and L. Dai, "External Text Based Data Augmentation for Low-Resource Speech Recognition in the Constrained Condition of OpenASR21 Challenge," in *Interspeech 2022*. ISCA, Sep. 2022, pp. 4860–4864.

[30] J. Zhao, H. Wang, J. Li, S. Chai, G. Wang, G. Chen, and W.-Q. Zhang, "The THUEE System Description for the IARPA OpenASR21 Challenge," in *Interspeech 2022*. ISCA, Sep. 2022, pp. 4855–4859.

[31] G. Huang, T. F. da Silva, L. Lamel, J.-L. Gauvain, A. Gorin, A. Laurent, R. Lileikyte, and A. Messouadi, "An investigation into language model data augmentation for low-resourced STT and KWS," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 5790–5794.