INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Unsupervised Cross-lingual Representation Learning for Speech Recognition

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, Michael Auli

Facebook AI, USA

aconneau@fb.com,abaevski@fb.com,locronan@fb.com@fb.com,abdo@fb.com,michaelauli@fb.com

Abstract

This paper presents XLSR which learns cross-lingual speech representations by pretraining a single model from the raw waveform of speech in multiple languages. We build on wav2vec 2.0 which is trained by solving a contrastive task over masked latent speech representations and jointly learns a quantization of the latents shared across languages. The resulting model is fine-tuned on labeled data and experiments show that cross-lingual pretraining significantly outperforms monolingual pretraining. On the CommonVoice benchmark, XLSR shows a relative phoneme error rate reduction of 72% compared to the best known results. On BABEL, our approach improves word error rate by 16% relative compared to a comparable system. Our approach enables a single multilingual speech recognition model which is competitive to strong individual models. We hope to catalyze research in low-resource speech understanding by releasing XLSR-53, a large model pretrained in 53 languages.

Index Terms: speech recognition, self-supervised learning, multilinguality, low-resource languages

1. Introduction

Cross-lingual learning aims to build models which leverage data from other languages to improve performance. This has been a long standing interest in the speech community [1, 2, 3, 4, 5, 6, 7] which includes systems able to transcribe multiple languages [8, 9, 10, 11, 12]. However, the vast majority of work in speech processing has focused on supervised cross-lingual training which requires labeled data in multiple languages. Transcribed speech is often much scarcer than unlabeled speech and requires non-trivial human annotation.

Unsupervised representation learning, or pretraining, does not require labeled data and has received a lot of recent attention in computer vision [13, 14] after much success in natural language processing [15]. For the latter, cross-lingual pretraining has been shown to be very effective, particularly, for low resource languages [16, 17]. In speech processing, most work in this area has focused on monolingual unsupervised representation learning [18, 19, 20, 21, 22, 23, 24, 25].

In this paper, we focus on the cross-lingual setting by learning representations on unlabeled data that generalize across languages. We build on the pretraining approach of [26] which jointly learns contextualized speech representations as well as a discrete vocabulary of latent speech representations. The latter serves to effectively train the model with a contrastive loss (§ 2) and the discrete speech representations are shared across languages. Different to recent work on unsupervised cross-lingual pretraining, we fine-tune the Transformer part of the model instead of freezing all pretrained representations [27] or feeding them to a separate downstream model [28]. We extend the work of [27] by pretraining on multiple languages instead of just English and we experiment on top of a stronger baseline.

We evaluate XLSR on 5 languages of the BABEL benchmark [29] which is conversational telephone data and ten languages of CommonVoice [30], a corpus of read speech (§ 3). Multilingual pretraining outperforms monolingual pretraining in most cases, except for resource rich languages and we show that increased model capacity significantly closes the gap. We also demonstrate that XLSR representations can be fine-tuned simultaneously on multiple languages to obtain a multilingual speech recognition system whose performance is competitive to fine-tuning a separate model on each language § 4).

2. Approach

Unsupervised cross-lingual representation learning has shown great success by pretraining Transformers with multilingual masked language models [15, 16]. In this work, we learn cross-lingual speech representations by extending wav2vec 2.0 [26] to the cross-lingual setting. Our approach learns a single set of quantized latent speech representations which are shared across languages. Next, we outline the architecture (§ 2.1), training (§ 2.2) and adaptations for cross-lingual training.

2.1. Architecture

We follow the design choices described in [26]. The model contains a convolutional feature encoder $f : \mathcal{X} \mapsto \mathcal{Z}$ to map raw audio $\mathcal{X}$ to latent speech representations $\mathbf{z}_1, \ldots, \mathbf{z}_T$ which are fed to a Transformer network $g : \mathcal{Z} \mapsto \mathcal{C}$ to output context representations $\mathbf{c}_1, \ldots, \mathbf{c}_T$ [15, 22, 25]. For the purpose of training the model, feature encoder representations are discretized to $\mathbf{q}_1, \ldots, \mathbf{q}_T$ with a quantization module $\mathcal{Z} \mapsto \mathcal{Q}$ to represent the targets in the self-supervised learning objective (§ 2.2). A Gumbel softmax enables choosing discrete codebook entries in a fully differentiable way [31]. Each $\mathbf{z}_t$ represents about 25ms of audio strided by 20ms, the context network architecture follows BERT [15] except for relative positional embeddings [25].

2.2. Training

The model is trained by solving a contrastive task over masked feature encoder outputs. For masking, we sample $p = 0.065$ of all time steps to be starting indices and mask the subsequent $M = 10$ time steps. The objective requires identifying the true quantized latent $\mathbf{q}_t$ for a masked time-step within a set of $K = 100$ distractors $\mathbf{Q}_t$ sampled from other masked time steps: $-\log \frac{\exp(sim(\mathbf{c}_t,\mathbf{q}_t))}{\sum_{\tilde{\mathbf{q}}\sim\mathbf{Q}_t}\exp(sim(\mathbf{c}_t,\tilde{\mathbf{q}}))}$ where $\mathbf{c}_t$ is the output of the transformer, and $sim(\mathbf{a}, \mathbf{b})$ denotes cosine similarity.

This is augmented by a codebook diversity penalty to encourage the model to use all codebook entries [32]. We maximize the entropy of the averaged softmax distribution over the codebook entries for each group $\bar{p}_g$ across a batch of utterances: $\frac{1}{GV}\sum_{g=1}^{G} -H(\bar{p}_g) = \frac{1}{GV}\sum_{g=1}^{G}\sum_{v=1}^{V}\bar{p}_{g,v}\log\bar{p}_{g,v}$. To stabilize the feature encoder we apply an L2 penalty over the outputs of the feature encoder.

When pretraining on L languages, we form multilingual batches [15, 16] by sampling speech samples from a multino-

http://dx.doi.org/10.21437/Interspeech.2021-329

mial distribution $(p_l)_{l=1,\ldots,L}$ where $p_l \sim \left(\frac{n_l}{N}\right)^\alpha$, n_l being the number of pretraining hours of language l, N the total number of hours, and α the upsampling factor. The parameter α controls the importance given to high-resource versus low-resource languages during pretraining.

3. Experimental setup

3.1. Datasets

CommonVoice. The CommonVoice dataset is a multilingual corpus of read speech comprising more than two thousand hours of speech in 38 languages [30]. The amount of data per language ranges from three hours for Swedish ("low-resource") to 353 hours for French and 1350 hours for English ("high-resource"). Following [27] we consider ten languages: *Spanish (es), French (fr), Italian (it), Kyrgyz (ky), Dutch (du), Russian (ru), Swedish (sv), Turkish (tr), Tatar (tt) and Chinese (zh)*; as well as English (en) for pretraining. We use the November 2019 release for training models, and for fine-tuning we use the evaluation splits of [27] which include one hour labeled data for training, 20 minutes for validation and one hour for testing. This few-shot evaluation dataset consists of phoneme sequences as output and we report phone error rate (PER) similar to prior work.

BABEL. This dataset is a multilingual corpus of conversational telephone speech from IARPA, which includes Asian and African languages [29]. We adopt the same setup as [6] and pretrain on ten languages: *Bengali (bn), Cantonese (zh), Georgian (ka), Haitian (ht), Kurmanji (ku), Pashto (ps), Tamil (ta), Turkish (tr), Tokpisin (tp), Vietnamese (vi).* We evaluate cross-lingual transfer on four other languages (models are not pretrained on these languages): *Assamese (as), Tagalog (tl), Swahili (sw), Lao (lo).* We train a multilingual model in ten languages and monolingual models in 14 languages. We use the same speech audio for pretraining and fine-tuning, and no unlabeled speech provided by BABEL. We use the dev folder of the BABEL dataset as our test set as "eval" has not been open-sourced, and use 10% of the training set as dev data. We report character error rate (CER) and word error rate (WER). All audio is resampled to 16kHz. For comparison with [33] only, we train 4-gram n-gram language models on CommonCrawl data [34, 35] for Assamese (140MiB of text data), Swahili (2GiB), Tamil (4.8GiB) and Lao (763MiB).

Multilingual LibriSpeech (MLS). The Multilingual LibriSpeech dataset [36] is a large corpus derived from read audiobooks of Librivox and consists of 8 languages: *Dutch (du), English (en), French (fr), German (de), Italian (it), Polish (pl), Portuguese (pt), Spanish (es).* The latest version of this corpus contains around 50k hours including 44k hours in English. We combine this corpus to CommonVoice and BABEL to train XLSR-53, a large pretrained model covering 53 languages.

3.2. Training details

Pretraining. Models are implemented in fairseq [37]. We evaluate two architectures with the same feature encoder (§ 2.1) but different Transformer settings: Base with 12 blocks, model dimension 768, inner dimension (FFN) 3072 and 8 attention heads; and Large with 24 blocks, model dimension 1024, inner dimension 4096 and 16 attention heads; both use dropout 0.1. For Base, we crop 250k samples, or 15.6sec of audio, and pack up to 1.4m samples on each GPU. For Large, we crop 320k samples and put up to 1.2m samples on a GPU. Batches are sampled using a factor $\alpha \in \{0.5, 1\}$. We use 16 GPUs for small datasets (typically monolingual) and 64 GPUs for large datasets (typically multilingual), and use Adam [38] where the learning rate is warmed up for the first 10% of updates to a peak of 1e-5 (Base) or 1e-3 (Large), and then linearly decayed over a total of 250k updates.

Fine-tuning. To fine-tune the model we add a classifier representing the output vocabulary of the respective downstream task on top of the model and train on the labeled data with a Connectionist Temporal Classification (CTC) loss [39, 25]. Weights of the feature encoder are not updated at fine-tuning time. We determine the best learning rates setting in [2e-5, 6e-5] based on dev set error rate. The learning rate schedule has three phases: warm up for the first 10% of updates, keep constant for 40% and then linearly decay for the remainder. For CommonVoice we fine-tune for 20k updates and on BABEL for 50k updates on 2 GPUs for the Base model and 4 GPUs for the Large model.

3.3. Pretrained models

We use the Base architecture unless otherwise stated. For CommonVoice, we pretrain an English model on 1350h, and ten monolingual models on each pretraining set. For comparison with the English model, we train Base and Large multilingual models on 1350h of data: 793h of speech audio from the 10 evaluation languages plus 557h of English audio. We upsample low-resource languages with $\alpha = 0.5$ and train a model with $\alpha = 1$ for comparison (unbalanced). For multilingual fine-tuning, we either separate or share phoneme vocabularies across languages. For BABEL, we train a monolingual model on each of the 14 languages, as well as a Base and Large multilingual model on a total of 650 hours of speech audio in ten languages. Since the amount of data in each language is more balanced than for CommonVoice, we use $\alpha = 1$. The same speech audio is used for pretraining and fine-tuning and we use separate character sets for multilingual fine-tuning. We train a large model on the 53 languages of MLS, CommonVoice and BABEL, which consists of 56 thousand hours of speech data. We call this model XLSR-53 and make it publicly available.

4. Results

In our experiments, we first show that our approach is very effective for learning generic cross-lingual representations in an unsupervised way. Pretraining a single model on multiple languages significantly outperforms the previous state of the art on CommonVoice, as well as our own monolingual models. Second, we demonstrate the positive impact of cross-lingual transfer on low-resource languages and provide a better understanding of the trade-off between high-resource and low-resource languages. Finally, by fine-tuning a multilingual model on many languages at once, we show that we can obtain a single model for all languages with strong performance.

4.1. Unsupervised cross-lingual representation learning

In what follows, we compare XLSR to several baselines and show that unsupervised cross-lingual representation learning is very effective. We provide a comprehensive analysis of the impact of different pretraining methods on automatic speech recognition in Table 1 and 2.

4.1.1. Multilingual outperforms mono pretraining and prior art

We first compare monolingual (XLSR-Monolingual) to multilingual (XLSR-10) pretrained models (Base) fine-tuned individually on each language (ft=1). On CommonVoice, XLSR-10 obtains 13.6 PER on average (*Avg*), a relative PER reduction of

Table 1: *CommonVoice phoneme error rate (PER). Models are pretrained on either one language (pt = 1) or 10 languages (pt = 10); and fine-tuned on each language (ft = 1) or all languages (ft = 10). D indicates the pretraining data, LS for English LibriSpeech (100h or 360h), BBL_{all} for BABEL (1070h), CV_{En} for English CommonVoice (1350h), CV_{mo} for monolingual (see number of pretraining hours per language) and CV_{all} for multilingual (1350h). Languages can be high-resource (es, fr, it) or low-resource (e.g. ky, sv, tr, tt).*

Model	**D**	**#pt**	**#ft**	**es**	**fr**	**it**	**ky**	**nl**	**ru**	**sv**	**tr**	**tt**	**zh**	**Avg**
Number of pretraining hours per language				168h	353h	90h	17h	29h	55h	3h	11h	17h	50h	793h
Number of fine-tuning hours per language				1h	1h	1h	1h	1h	1h	1h	1h	1h	1h	10h
Baselines from previous work														
m-CPC† [27]	LS_{100h}	10	1	38.7	49.3	42.1	40.7	44.4	45.2	48.8	49.7	44.0	55.5	45.8
m-CPC† [27]	LS_{360h}	10	1	38.0	47.1	40.5	41.2	42.5	43.7	47.5	47.3	42.0	55.0	44.5
Fer et al.† [40]	BBL_{all}	10	1	36.6	48.3	39.0	38.7	47.9	45.2	52.6	43.4	42.5	54.3	44.9
Our monolingual models														
XLSR-English	CV_{en}	1	1	13.7	20.0	19.1	13.2	19.4	18.6	21.1	15.5	11.5	27.1	17.9
XLSR-Monolingual	CV_{mo}	1	1	6.8	10.4	10.9	29.6	37.4	11.6	63.6	44.0	21.4	31.4	26.7
Our multilingual models														
XLSR-10 (unbalanced)	CV_{all}	10	1	9.7	13.6	15.2	11.1	18.1	13.7	21.4	14.2	9.7	25.8	15.3
XLSR-10	CV_{all}	10	1	9.4	14.2	14.1	8.4	16.1	11.0	20.7	11.2	7.6	24.0	13.6
XLSR-10	CV_{all}	10	10	10.0	13.8	14.0	8.8	16.5	11.6	21.4	12.0	8.7	24.5	14.1
Our multilingual models (Large)														
XLSR-10	CV_{all}	10	1	7.9	12.6	11.7	7.0	14.0	9.3	20.6	9.7	7.2	22.8	12.3
XLSR-10	CV_{all}	10	10	8.1	12.1	11.9	7.1	13.9	9.8	21.0	10.4	7.6	22.3	12.4
Our Large XLSR-53 model pretrained on 56k hours														
XLSR-53	D_{53}	53	1	2.9	5.0	5.7	6.1	5.8	8.1	12.2	7.1	5.1	18.3	7.6

49% compared to XLSR-Monolingual (Table 1). On BABEL, XLSR-10 improves over XLSR-Monolingual by 22% relative WER (Table 2) and by more over supervised training (Training from scratch). Pretraining on multiple languages results in cross-lingual transfer and better speech representations.

Compared to prior work, XLSR-10 Large reduces PER by 72% relative to m-CPC [27] on CommonVoice (Table 1). For BABEL, the most comparable work is [33] and in Table 3, XLSR-10 shows a relative word error reduction of 16% compared to their BLSTM-HMM baseline. In Georgian, we also see that our XLSR-53 model, which leverages additional unsupervised data from other datasets, obtains 31.1 which outperforms the best system by [41]. Our approach is also not tuned to the limit, for instance we do not consider Transformer language models but only a 4-gram KenLM.

4.1.2. Multilingual outperforms English-only pretraining

To isolate the impact of multilingual training versus simply training on more data, we pretrain an English-only CommonVoice model (XLSR-English) on the same amount of data as the multilingual model (1350h) and compare the two. Table 1 shows that on average, XLSR-English significantly improves over the monolingual models (average PER of 26.7 vs. 17.9 PER) but multilingual pretraining performs even better at 13.6 PER, a 24% relative PER reduction over XLSR-English. This shows that adding more training data is not the only reason for the improved accuracy: the similarity between the languages used in pretraining and fine-tuning also plays an important role. To enable few-shot learning with self-supervised learning on languages that are more low-resource, it is thus important to pretrain in multiple languages at once.

4.1.3. XLSR representations transfer well to unseen languages

To better assess the cross-lingual transfer of the learned representations, we evaluate the XLSR-10 BABEL model on four languages not seen during pretraining. We fine-tune this model on each language, and compare it to monolingual models pretrained specifically on these languages. Table 2 shows that a multilingual model not pretrained on any data from the four languages, still outperforms XLSR-Monolingual, reducing average CER from 29 to 22.8 which compares to results from previous work of 36.8 CER [6]. This further suggests that the learned representations capture generic features of the speech signal which transfer to many languages.

4.2. Understanding cross-lingual transfer learning

In this section, we examine several properties of unsupervised cross-lingual representation learning for speech recognition. We show that it is particularly effective on low-resource languages, then describe the transfer-interference trade-off which benefits low resource languages but hurts high resource languages. Finally, we show that adding capacity is important for multilingual pretraining.

4.2.1. Cross-lingual transfer learning improves low-resource language understanding

Unsupervised cross-lingual representation learning and cross-lingual transfer are particularly effective on low-resource languages. On CommonVoice, the separation between high and low-resource languages is more salient than for BABEL. We distinguish between low-resource and high-resource based on the amount of available unlabeled speech data. For example, French and Spanish have 353h and 168h and are thus high-resource, while Swedish and Turkish have 3h and 11h and are low-resource.

Table 2: *BABEL results on out-of-pretraining languages (CER). XLSR-10 provides strong representations for languages not seen during pretraining, outperforming monolingual models pretrained specifically on these languages.*

Model	#pt	#ft	as	tl	sw	lo	Avg
Number of pretraining hours			55h	76h	30h	59h	220h
Number of fine-tuning hours			55h	76h	30h	59h	220h
Baselines from previous work							
Cho et al. [6] (Mono)	10	1	45.6	43.1	33.1	42.1	41.0
Cho et al. [6] (Stage-2)	10	1	41.3	37.9	29.1	38.7	36.8
Our monolingual models							
Training from scratch	1	1	50.2	41.7	40.8	43.5	44.1
XLSR-Monolingual	1	1	34.8	25.4	26.8	29.1	29.0
Our multilingual models							
XLSR-10	10	1	29.4	21.9	16.6	23.3	22.8
XLSR-10 (Large)	10	1	27.7	19.6	14.9	21.8	21.0
XLSR-53 (Large)	53	1	17.9	13.1	21.3	22.4	18.7

Table 3: *BABEL results on out-of-pretraining languages using word error rate (WER). We report baselines from previous work, including the strong [41] baseline. We report WER with and without 4-gram KenLM.*

Model	#pt	#ft	as	tl	sw	lo	ka
Number of pretraining hours			55h	76h	30h	59h	46h
Number of fine-tuning hours			55h	76h	30h	59h	46h
Baselines from previous work							
Alumae et al. [41]	1	1	-	-	-	-	32.2
Ragni et al. [42]	1	1	-	40.6	35.5	-	-
Inaguma et al. [33]	1	1	49.1	46.3	38.3	45.7	-
Our approach (no LM)							
XLSR-10 (Large)	10	1	49.1	40.6	38.1	34.7	-
Our approach (4-gram KenLM)							
XLSR-10 (Large)	10	1	44.9	37.3	35.5	32.2	-
XLSR-53 (Large)	53	1	44.1	33.2	26.5	-	31.1

Monolingual models perform poorly on low-resource languages but this is where cross-lingual transfer is most effective: XLSR-10 reduces PER over XLSR-Monolingual by a relative 67% on Swedish, 72% on Turkish, 72% on Kyrgyz, and 64% on Tatar.

On BABEL, results in Table 2 show that XLSR-10 outperforms XLSR-Monolingual on all languages. When evaluating on all BABEL languages, we observed that the biggest gains for XLSR-10 were obtained on the four lowest-resource languages: Georgian (ka), Kurmanji (ku), Tokpisin (tp) and Swahili (sw).

We also observe in Table 1 strong cross-domain transfer from the MLS data to the CommonVoice data which are both read-speech. The XLSR-53 increases performance particularly on languages that are shared between the two datasets. For example, in Spanish and French XLSR-53 outperforms XLSR-10 by 4.8 and 5.9% absolute PER. This model also performs well on other languages, including BABEL (see Table 2) on which it reduces word error rate by 2.3% WER over XLSR-10.

4.2.2. *The transfer-interference trade-off: high vs. low-resource*

Per language results on CommonVoice (Table 1) show *transfer-interference* trade-off [43]: for low-resource languages (e.g. ky, nl, sv, tr, tt), multilingual models outperform monolingual models because of positive transfer, however multilingual models perform worse on high-resource languages (es, fr, it), due to *interference*. Data from multiple languages enables better speech representations that transfer to low-resource languages but the model also needs to share its capacity across languages which degrades performance on high-resource languages.

For a given model capacity, the language sampling parameter α (see § 2) controls this trade-off. Table 1 shows that training according to the true language distribution, XLSR-10 (unbalanced) using $\alpha = 1$, performs less well than XLSR-10, where more capacity is allocated to low-resource languages via $\alpha = 0.5$. The sole exception being French, the language with the most data. On average the unbalanced model obtains 15.3 PER while the balanced model obtains 13.6.

4.2.3. *Increasing capacity for a multilingual pretrained model*

The interference problem can be alleviated by adding more capacity to the multilingual model [43, 17]: the gap between multilingual models and monolingual models for high-resource languages can be reduced by increasing model capacity. In this work, we only study the impact of adding more capacity to the multilingual model, by training an XLSR-10 Large model. On CommonVoice, the Large model reduces PER by relative 9.6% compared to Base, reducing average PER from 13.6 to 12.3. There are no gains on very low-resource languages like Swedish but significant gains on Spanish, French and Italian. On BABEL, average CER is reduced by a relative 6.8%. This shows that the multilingual model benefits from more capacity overall, and in particular for high-resource languages.

4.3. One model for all languages (multilingual fine-tuning)

When we fine-tune the pretrained model on each language individually, then we end up with a different model for each language. On the other hand, multilingual speech recognition aims to build a single model for all languages that performs as well or better than individual monolingual models. Next, we investigate fine-tuning a single model on the labeled data of all languages (#ft=10) to obtain a single multilingual model instead of fine-tuning each language separately (#ft=1). Training batches are constructed by sampling audio samples from multiple languages (without upsampling).

For CommonVoice Table 1 shows that the Base model with monolingual fine-tuning of XLSR-10 obtains 13.6 average PER which compares to 14.1 PER for multilingual fine-tuning. When increasing model capacity (Large), multilingual fine-tuning is competitive to monolingual fine-tuning: 12.3 average PER (ft=1) vs. 12.4 average PER (ft=N). Increasing capacity is particularly important when fine-tuning on large amounts of supervised data from many languages. Multilingual fine-tuning performs competitively to monolingual fine-tuning and enables us to have a single model for many languages.

5. Conclusion

In this work, we investigated unsupervised cross-lingual speech representations learned from the raw waveform. We show that pretraining on data in many languages improves both over monolingual pretraining and prior work, with the largest improvements on low-resource languages. Fine-tuning the model on multiple languages at once enables a single multilingual speech recognition model competitive to individually fine-tuned models.

6. References

[1] W. Byrne, P. Beyerlein, J. M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, D. Vergyri, and T. Wang, "Towards language independent acoustic modeling," in *ICASSP*, 2000.

[2] V.-B. Le and L. Besacier, "Automatic speech recognition for under-resourced languages: Application to vietnamese language," *IEEE Transactions on Audio, Speech, and Language Processing*, 2009.

[3] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *ICASSP*, 2013.

[4] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *ICASSP*, 2013.

[5] M. J. Gales, K. M. Knill, and A. Ragni, "Low-resource speech recognition and keyword-spotting," in *SPECOM*, 2017.

[6] J. Cho, M. K. Baskar, R. Li, M. Wiesner, S. H. Mallidi, N. Yalta, M. Karafiát, S. Watanabe, and T. Hori, "Multilingual sequence-to-sequence speech recognition: Architecture, transfer learning, and language modeling," in *IEEE SLT*, 2018.

[7] H. Seki, S. Watanabe, T. Hori, J. L. Roux, and J. R. Hershey, "An end-to-end language-tracking speech recognizer for mixed-language speech," in *ICASSP*, 2018.

[8] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, D. Povey, A. Rastrow, R. Rose, and S. Thomas, "Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models," in *ICASSP*, 2010.

[9] H. Bourlard, J. Dines, M. Magimai-Doss, P. Garner, D. Imseng, P. Motlicek, H. Liang, L. Saheer, and F. Valente, "Current trends in multilingual speech processing," *Sadhana*, vol. 36, 2011.

[10] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *ICASSP*, 2013.

[11] 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 *ICASSP*, 2018.

[12] 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.

[13] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," *arXiv*, vol. abs/1911.05722, 2019.

[14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv*, vol. abs/2002.05709, 2020.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv*, vol. abs/1810.04805, 2018.

[16] G. Lample and A. Conneau, "Cross-lingual language model pre-training," in *NeurIPS*, 2019.

[17] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv*, vol. abs/1911.02116, 2019.

[18] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv*, vol. abs/1807.03748, 2018.

[19] Y.-A. Chung and J. Glass, "Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech," *arXiv*, vol. abs/1803.08976, 2018.

[20] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Interspeech*, 2019.

[21] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," *arXiv*, vol. abs/1904.03240, 2019.

[22] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *ICLR*, 2020.

[23] D. Harwath, W.-N. Hsu, and J. Glass, "Learning hierarchical discrete linguistic units from visually-grounded speech," in *ICLR*, 2020.

[24] R. Eloff, A. Nortje, B. van Niekerk, A. Govender, L. Nortje, A. Pretorius, E. Van Biljon, E. van der Westhuizen, L. van Staden, and H. Kamper, "Unsupervised acoustic unit discovery for speech synthesis using discrete latent-variable neural networks," *arXiv*, vol. abs/1904.07556, 2019.

[25] A. Baevski, M. Auli, and A. Mohamed, "Effectiveness of self-supervised pre-training for speech recognition," in *ICASSP*, 2020.

[26] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv*, vol. abs/2006.11477, 2020.

[27] M. Rivière, A. Joulin, P.-E. Mazaré, and E. Dupoux, "Unsupervised pretraining transfers well across languages," *arXiv*, vol. abs/2002.02848, 2020.

[28] K. Kawakami, L. Wang, C. Dyer, P. Blunsom, and A. van den Oord, "Learning robust and multilingual speech representations," *arXiv*, vol. abs/2001.11128, 2020.

[29] M. J. F. Gales, K. M. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low-resource languages: Babel project research at cued," in *Spoken Language Technologies for Under-Resourced Languages*, 2014.

[30] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv*, vol. abs/1912.06670, 2019.

[31] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv*, vol. abs/1611.01144, 2016.

[32] S. Dieleman, A. van den Oord, and K. Simonyan, "The challenge of realistic music generation: modelling raw audio at scale," *arXiv*, 2018.

[33] H. Inaguma, J. Cho, M. K. Baskar, T. Kawahara, and S. Watanabe, "Transfer learning of language-independent end-to-end asr with language model fusion," in *ICASSP*, 2019.

[34] K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn, "Scalable modified Kneser-Ney language model estimation," in *ACL*, 2013.

[35] G. Wenzek, M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and E. Grave, "Ccnet: Extracting high quality monolingual datasets from web crawl data," 2019.

[36] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," *Interspeech 2020*, pp. 2757–2761, 2020.

[37] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *NAACL System Demonstrations*, 2019.

[38] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *ICLR*, 2015.

[39] A. Graves, S. Fernández, and F. Gomez, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006.

[40] R. Fer, P. Matějka, F. Grézl, O. Plchot, K. Veselỳ, and J. H. Černockỳ, "Multilingually trained bottleneck features in spoken language recognition," *Computer Speech & Language*, vol. 46, pp. 252–267, 2017.

[41] T. Alumäe, D. Karakos, W. Hartmann, R. Hsiao, L. Zhang, L. Nguyen, S. Tsakalidis, and R. Schwartz, "The 2016 bbn georgian telephone speech keyword spotting system," in *ICASSP*, 2017.

[42] A. Ragni, Q. Li, M. J. F. Gales, and Y. Wang, "Confidence estimation and deletion prediction using bidirectional recurrent neural networks," in *SLT*, Athens, 2018.

[43] N. Arivazhagan, A. Bapna, O. Firat, D. Lepikhin, M. Johnson, M. Krikun, M. X. Chen, Y. Cao, G. Foster, C. Cherry *et al.*, "Massively multilingual neural machine translation in the wild: Findings and challenges," *arXiv preprint arXiv:1907.05019*, 2019.