Learning Robust and Multilingual Speech Representations

Kazuya Kawakami[♣♠] **Luyu Wang**[♣] **Chris Dyer**[♣] **Phil Blunsom**[♣♠] **Aaron van den Oord**[♣]
[♣]DeepMind, London, UK
[♠]Department of Computer Science, University of Oxford, Oxford, UK
{kawakamik,luyuwang,cdyer,pblunsom,avdnoord}@google.com

Abstract

Unsupervised speech representation learning has shown remarkable success at finding representations that correlate with phonetic structures and improve downstream speech recognition performance. However, most research has been focused on evaluating the representations in terms of their ability to improve the performance of speech recognition systems on read English (e.g. Wall Street Journal and LibriSpeech). This evaluation methodology overlooks two important desiderata that speech representations should have: robustness to domain shifts and transferability to other languages. In this paper we learn representations from up to 8000 hours of diverse and noisy speech data and evaluate the representations by looking at their robustness to domain shifts and their ability to improve recognition performance in many languages. We find that our representations confer significant robustness advantages to the resulting recognition systems: we see significant improvements in out-of-domain transfer relative to baseline feature sets and the features likewise provide improvements in 25 phonetically diverse languages.

1 Introduction

The input representation of machine learning model strongly determines the difficulty faced by the learning algorithm, how much data the learner will require to find a good solution, and whether the learner generalizes out of sample and out of the domain of the training data. Representations (or features) that encode relevant information about data enable models to achieve good performance on downstream tasks, while representations that are invariant to factors that are not relevant to downstream tasks can further improve generalization. Traditionally, many invariances were hard-coded in feature extraction methods. For example, in image representations, geometric and photometric invariance has been investigated (Mundy et al., 1992; Van De Weijer et al., 2005). For acoustic representations, standard MFCC features are sensitive to additive noise and many modifications have been proposed to overcome those limitations (Dev and Bansal, 2010; Kumar et al., 2011).

Recently, unsupervised representation learning algorithms have shown significant improvements at learning representations that correlate well with phonetic structure (van den Oord et al., 2018; Kahn et al., 2019b) and improving downstream speech recognition performance (Schneider et al., 2019; Baevski et al., 2019). Most of this work focused on learning representations from read English speech (from the LibriSpeech and LibriVox datasets) and evaluating the features when used to recognize speech in a rather similar domain (read English text). However, this approach to evaluation fails to test for the invariances that we would like good speech representations to have: robustness to domain shifts and transferability to other languages.

In our experiments we learn representations from 8000 hours of diverse and noisy speech, using an extended version of contrastive predictive coding model: bidirectional predictive models with dense residual connections (§2–§4), and evaluate the robustness and transferability of our representations by estimating how invariant they are to domain and language shifts. To do so, an ASR model is trained using our representations on one dataset but evaluated on the test sets of other datasets. In this experiment, we find that the representations derived from the large pretraining dataset lead the ASR model to be much more robust to domain shifts, compared to both log filterbank features as well as to pretraining just on LibriSpeech. We also train ASR models on 25 languages, including low-resource languages (e.g. Amharic, Fongbe, Swahili, Wolof), and show that our representations significantly outperform

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1182–1192

both standard features and those pretrained only on clean English data in the language transfer setup.

In summary, we confirm several increasingly common patterns that may be discerned in the literature on unsupervised representation learning, across a variety of modalities. First, scale matters: good representation learning requires a large amount of data. Second, unsupervised representations consistently improve robustness on downstream tasks. And finally, representations learned from multilingual data can transfer across many languages.

2 Contrastive Predictive Coding: CPC

Unsupervised representation learning methods rely on differentiable objectives which quantify the degree to which representations have succeeded at capturing the relevant characteristics in data. Mutual information measures relationships between random variables (Fano and Hawkins, 1961). Mutual information maximization techniques, that learn representations that describe data by maximizing mutual information between data and representation variables, have been explored for a long time in unsupervised representation learning (Linsker, 1988; Bell and Sejnowski, 1995). However, since the exact computation of mutual information is not tractable for continuous variables, recently many estimators have been proposed for enabling unsupervised representation learning with neural networks (Belghazi et al., 2018; van den Oord et al., 2018; Poole et al., 2019).

Contrastive predictive coding (van den Oord et al., 2018, CPC) is a mutual information maximization method that has been successfully applied to many modalities such as images and speech (Hénaff et al., 2019; Schneider et al., 2019). The objective is designed to extract features that allow the model to make long-term predictions about future observations. This is done by maximizing the mutual information of these features with those extracted from future timesteps. The intuition is that the representations capture different levels of structure dependent on how far ahead the model predicts. For example, if the model only predicts a few steps ahead, the resulting representations can capture local structures. On the other hand, if the model predicts further in the future, the representations will need to infer "slow features" (Wiskott and Sejnowski, 2002); more global structures such as phonemes, words and utterances in speech.

The overall unsupervised learning process is visualized in Figure 1. Given a raw audio signal of length L ($\boldsymbol{x} = x_1, x_2, \ldots, x_L$, $x_i \in \mathbb{R}$ where x_i represents the acoustic amplitude at time i), a function g_{enc} encodes the audio signals into vector representations ($\boldsymbol{z} = \boldsymbol{z}_1, \boldsymbol{z}_2 \ldots, \boldsymbol{z}_M$, $\boldsymbol{z} \in \mathbb{R}^{d_z}$). Next, an autoregressive function g_{ar}, such as a recurrent neural network, summarizes the past representations and produces context vectors ($\boldsymbol{c} = \boldsymbol{c}_1, \boldsymbol{c}_2 \ldots, \boldsymbol{c}_M, \boldsymbol{c} \in \mathbb{R}^{d_c}$). The representations are learned to maximize mutual information between context vectors ($\boldsymbol{c}_t$) and future latent representations ($\boldsymbol{z} + \boldsymbol{k}$) as follows:

$$I(\boldsymbol{c}_t, \boldsymbol{z}_{t+k}) = \sum_{\boldsymbol{c}_t, \boldsymbol{z}_{t+k}} p(\boldsymbol{c}_t, \boldsymbol{z}_{t+k} \mid k) \log \frac{p(\boldsymbol{z}_{t+k} \mid \boldsymbol{c}_t, k)}{p(\boldsymbol{z}_{t+k})}.$$

Since the mutual information is not tractable for high dimensional data, it is common to use a lower-bound on the mutual information such as InfoNCE (van den Oord et al., 2018) which is a loss function based on noise contrastive estimation (Gutmann and Hyvärinen, 2010). Given a set $Z = \{\boldsymbol{z}_1, \ldots \boldsymbol{z}_N\}$ which contains one positive sample from $p(\boldsymbol{z}_{t+k}|\boldsymbol{c}_t)$ and $N-1$ negative samples from a "noise" distribution $p(\boldsymbol{z})$, the approximated lower-bound is written as:

$$I(\boldsymbol{c}_t, \boldsymbol{z}_{t+k}) \geq \mathbb{E}_Z \left[\log \frac{f_k(\boldsymbol{c}_t, \boldsymbol{z}_{t+k})}{\frac{1}{N} \sum_{\tilde{\boldsymbol{z}} \in Z} f_k(\boldsymbol{c}_t, \tilde{\boldsymbol{z}})} \right] = \mathcal{L}_{tk}^{NCE},$$

where $f_k(\boldsymbol{c}_t, \boldsymbol{z}_{t+k})$ is a scoring function. We used the standard log-bilinear model as follows:

$$f_k(\boldsymbol{c}_t, \boldsymbol{z}_{t+k}) = \exp(\boldsymbol{c}_t^T \boldsymbol{W}_k \boldsymbol{z}_{t+k}).$$

The loss function we maximize is a sum of the InfoNCE loss for each step, $\mathcal{L}^{NCE} = \sum_t \sum_k \mathcal{L}_{tk}^{NCE}$ and the negatives are uniformly sampled from representations in the same audio signal ($\boldsymbol{z}$).

3 Methods

In this section, we describe our models and objectives for unsupervised representation learning and downstream speech recognition. First, an acoustic feature extractor is trained with a bidirectional variant of contrastive predictive coding on an unlabeled audio dataset. Next, the parameters of this model are frozen and its output representations are used as input to train various speech recognition models, potentially on a different or smaller labeled dataset (Figure 1).

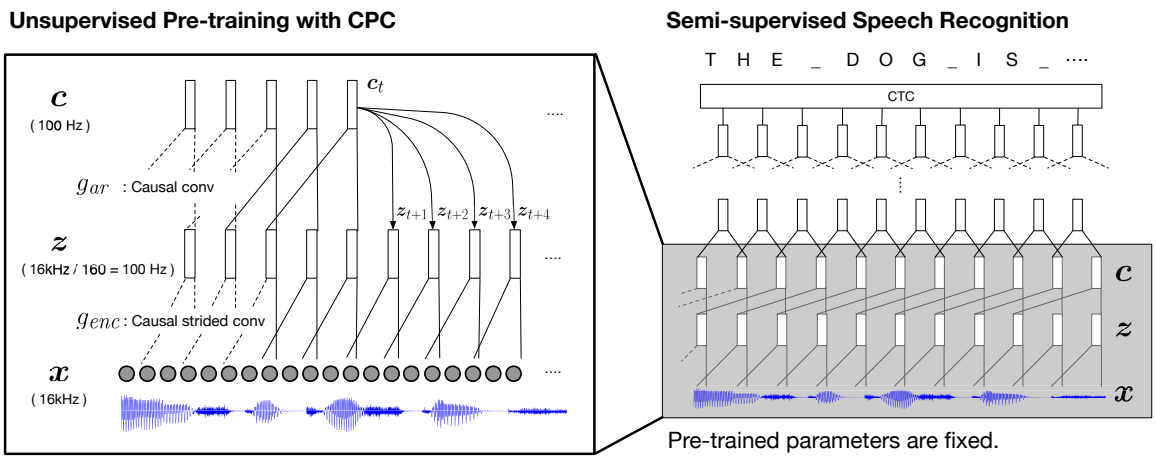

Figure 1: **Left**, unsupervised representation learning with forward contrastive predictive coding. The learned representations are fixed and used as inputs to a speech recognition model (**Right**).

3.1 Unsupervised learning with bi-directional CPC

Following the success of bidirectional models in representation learning (Peters et al., 2018; Devlin et al., 2019), we extend the original CPC method explained above with bidirectional context networks. The encoder function g_{enc} is shared for both directions, but there are two autoregressive models (g_{ar}^{fwd} and g_{ar}^{bwd}) which read encoded observations ($\boldsymbol{z}$) from the forward and backward contexts, respectively. The forward and backward context representations $\boldsymbol{c}_t^{fwd}, \boldsymbol{c}_t^{bwd}$ are learned with separate InfoNCE losses. When they are used for downstream tasks, a concatenation of two representations $\boldsymbol{c_t} = [\boldsymbol{c}_t^{fwd}; \boldsymbol{c}_t^{bwd}]$ is used. A similar technique has been used in image representation learning where representations are learned along different spatial dimensions (Hénaff et al., 2019).

All audio signals have a sampling rate of 16kHz and we normalize the mean and variance of the input signals over each utterance in order to mitigate volume differences between samples. For architectures, we use encoder and autoregressive models similar to (Schneider et al., 2019). The encoder function g_{enc}, is a stack of causal convolutions with kernel sizes (10, 8, 4, 4, 4, 1, 1) and stride sizes (5, 4, 2, 2, 2, 1, 1), corresponding to a receptive field of 10 ms of audio. For autoregressive functions, we use a 13 layer causal convolution architecture with kernel sizes (1, 2, . . . , 12, 13) and stride size 1, for both forward and backward functions. Layer-normalization across the temporal and feature dimensions is applied to every layer. Also, each layer has dense skip connections with layers below as in DenseNet (Huang et al., 2017). The objective function we optimize is the sum of the forward and backward InfoNCE losses (eq.2).

3.2 Semi-supervised speech recognition

Once the acoustic representations are trained, the resulting context vectors ($\boldsymbol{c}$) are used as inputs to character-level speech recognition models which predict transcriptions of audio-signals character by character. The model first predicts frame-level character probabilities with a series of convolution layers while the CTC forward algorithm (Graves et al., 2006) calculates conditional probabilities of a transcription given an audio signal. The model parameters are trained to maximize the log likelihood of the data. The training terminates when the word error rate on the development set stops improving or the model has trained for more than a certain number of epochs. The models are evaluated on the standard word error rate (WER) metric on held-out test data. During training, the parameters in the speech recognition models are trained with supervision but the parameters of the representation models remain fixed. For decoding, we use greedy CTC decoding. In most experiments, we do not use a language model (LM) in order to isolate the effects of the acoustic representations, but we do include results with a 4-gram LM to facilitate comparisons with published results.

Common practice in unsupervised representation learning is to evaluate learned representations using a linear classifier rather than a more complex non-linear model. However, we find that a simple linear layer followed by a CTC decoder does not have enough capacity to recognize speech. Thus, for our

first set of experiments we use a smaller version of DeepSpeech2 (Amodei et al., 2016) to predict the frame-level character probabilities. The model has two 2d-convolutions with kernel sizes (11, 41) and (11, 21) and stride sizes (2, 2) and (1, 2) and one unidirectional recurrent neural network (GRU) on top of the output from the convolution layers. A linear transformation and a softmax function are applied to predict frame-level character probabilities. We refer to **DeepSpeech2 small** for the model specifics (Amodei et al., 2016). In order to further investigate how the representations interact with larger speech recognition models, we use the time-delay neural networks (**TDNN**) that are commonly used in speech recognition (Collobert et al., 2016; Kuchaiev et al., 2018). These consist of 17 layers of 1d-convolutions followed by 2 fully connected layers. Refer to OpenSeq2Seq for a detailed description.[1] These large models have been designed to perform well with log-filterbank features and purely supervised learning on large datasets, so they represent a challenging and informative test case for the value of learned representations.

4 Experiments and Results

4.1 Datasets

We collected publicly available speech datasets which cover a variety of types of speech (e.g. read and spoken), noise conditions and languages. For unsupervised pretraining we use a combination of datasets, using the audio but not any transcriptions, even when they are available. For semi-supervised learning (i.e., evaluation) on top of the representations we use the transcribed datasets following their standard train-test splits. Table 1 summarizes the datasets used for unsupervised learning and English speech recognition tasks.

Unlabeled speech pretraining corpus For pretraining, we collected a diverse and noisy speech corpus from several existing datasets: the subset of Audio Set (Gemmeke et al., 2017) containing speech examples, the audio part of AVSpeech (Ephrat et al., 2018), and the Common Voice (CV)[2] dataset in all 29 available languages. In addition we used the audio from TIMIT (Garofolo, 1993) and the Speech Accent Archive (Weinberger and Kunath, 2009), ignoring the transcriptions. Finally, we include the audio (again ignoring transcriptions) from the standard training splits of the evaluation datasets below. This collection spans a range of recording conditions, noise levels, speaking styles, and languages and amounts to about 8000 hours of audio.

Name	Language	Type	Hours
Audio Set	Multilingual	-	2500
AVSpeech	Multilingual	-	3100
Common Voice	Multilingual	read	430
LibriSpeech	English	read	960
WSJ	English	read	80
TIMIT	English	read	5
SSA	English	read	<1
Tedlium	English	spoken	440
Switchboard	English	spoken	310

Table 1: Summary of English Datasets.

Transcribed read English For evaluation, we look at the performance of our representations on a variety of standard English recognition tasks, as well as their ability to be trained on one and tested on another. For read English, we use LibriSpeech (Panayotov et al., 2015) and the Wall Street Journal (Paul and Baker, 1992).

Transcribed spoken English To explore more extreme domain shifts, we additionally used conversational speech and public speaking datasets. We used Switchboard (Godfrey et al., 1992), a standard conversational speech recognition dataset consisting of two-sided telephone conversations (test only). Since the data was recorded more than 10 years ago and at a lower sampling rate than the other corpora, it presents a noisy and challenging recognition problem. Finally, we also use the Tedlium-3 (Hernandez et al., 2018) corpus, a large spoken English dataset containing 450 hours of speech extracted from TED conference talks. The recordings are clear, but there is some reverberation.

Transcription normalization Since we are comparing ASR systems trained on one dataset but evaluated on the test set of another, we normalize transcriptions to reduce systematic biases in the transfer condition. To do so, we use the format of the LibriSpeech dataset, which also ensures that our results are comparable with standard speech recognition systems on that task (Kuchaiev et al., 2018). For the other datasets, transcriptions are lowercased

[1] https://nvidia.github.io/OpenSeq2Seq/html/speech-recognition/wave2letter.html

[2] https://voice.mozilla.org

and unpronounced symbols (e.g., punctuation, silence markers) are removed. We also remove utterances containing numbers as they are transcribed inconsistently across and within datasets.

Transcribed multilingual speech In order to evaluate the transferability of the representations, we use speech recognition datasets in 4 African languages collected by the ALFFA project,[3] Amharic (Tachbelie et al., 2014), Fongbe (A. A Laleye et al., 2016), Swahili (Gelas et al., 2012), Wolof (Gauthier et al., 2016), for evaluation. These languages have unique phonological properties (e.g. height harmony) and phonetic inventories, making them a good contrast to English. These African languages are low-resource, each with 20 hours or less of transcribed speech. We also use 21 phonetically diverse languages from OpenSLR.[4] See Appendix A for more detail.

4.2 Unsupervised Representation Learning

We train the model described above (§3.1) using the datasets described in the previous section (§4.1). Similarly to Schneider et al. (2019)), audio signals are randomly cropped with a window size 149,600 observations (9.35 seconds) and encoded with the model. The bidirectional contrastive predictive coding objective (Eq. 2) with prediction steps (k) 12 and negatives (N) 10 is optimized with the Adam optimizer with learning rate 0.0001. A batch size of 128 is used as well as a polynomial learning rate scheduler with power 2 and gradient clipping with maximum norm 5.0. Training was terminated at 4.2 million steps based on speech recognition performance on the dev (= validation) set of the LibriSpeech corpus.

4.3 Robustness

Robustness to shifts in domain, recording conditions, and noise levels is an important desideratum for a good ASR system, and we hypothesized that the diversity of our largest pretraining regime would improve robustness along these dimensions. In contrast, standard MFCC features have been tested in terms of noise robustness and it is known that such representations are sensitive to additive noise (Zhao and Wang, 2013). Moreover, speech recognition systems developed on top of such features are not robust when they are evaluated on out-of-domain datasets (Amodei et al., 2016).

To test whether our pretraining approach improves robustness, we evaluate speech recognition models trained on the learned representations on many different datasets so as to investigate benefit of using the representations learned from large-scale data. We compare ASR systems on all of the Wall Street Journal and LibriSpeech corpora with the same optimization as explained above and evaluate word error rate on different evaluation sets, such as phone call conversations (Switchboard).

Table 2 summarizes the results on models trained on Wall Street Journal, LibriSpeech or the Tedlium corpora and evaluated on different evaluation sets. **CPC-LibriSpeech** and **CPC-8k** indicate representations are learned from LibriSpeech and 8000h of speech datasets listed above respectively. The features trained on large-scale data consistently outperform other representations across different evaluation sets. The speech recognition models trained on the Wall Street Journal perform badly on phone call data in general. However, CPC representations learned on large datasets are more robust than those trained only on read English data (LibriSpeech).

4.4 Low-resource Languages

Thus far, all our experiments have compared our representations in terms of their impacts on English recognition tasks (although we know that the pretraining dataset contains samples from many languages). We now turn to the question of whether these representations are suitable for driving recognition different languages with substantially different phonetic properties than English has. Specifically, we look at the performance on four languages—Amharic, Fongbe, Swahili, and Wolof—which manifest a variety of interesting phonological properties that are quite different from English. Evaluating on such languages will provide insights into the phonetic space learned in the representations. Moreover, our non-English languages are low-resource in terms of speech recognition data, but have 2–20 million native speakers each. It is therefore valuable if the representations learned from large-scale unlabelled data can improve low-resource speech recognition. Although there is a chance that the large-scale pretraining dataset may contain some examples from those languages, we did not add any extra data specifically from those languages.

To test the cross-linguistic value of these features, we trained speech recognition models on

[3] http://alffa.imag.fr
[4] https://openslr.org

	WSJ		LibriSpeech		Tedlium		Switchboard
	test92	test93	test-clean	test-other	dev	test	eval2000
WSJ							
LogFilterbank	16.78	23.26	46.27	73.27	58.61	62.55	96.44
CPC-LibriSpeech	11.89	15.66	31.05	56.31	45.42	47.79	83.08
CPC-8k	**10.77**	**14.99**	**29.18**	**51.29**	**38.46**	**39.54**	**69.13**
LibriSpeech							
LogFilterbank	14.42	21.08	6.43	20.16	26.9	25.94	61.56
CPC-LibriSpeech	14.28	20.74	6.91	21.6	26.53	27.14	63.69
CPC-8k	**13.31**	**18.88**	**6.25**	**19.10**	**21.56**	**21.77**	**53.02**
Tedlium							
LogFilterbank	20.35	27.23	24.05	47.27	18.75	19.31	74.55
CPC-LibriSpeech	15.01	19.52	17.77	36.7	15.28	15.87	61.94
CPC-8k	**13.17**	**17.75**	**16.03**	**32.35**	**13.67**	**13.88**	**47.69**

Table 2: Domain transfer experiments to test the robustness of the representations to domain shifts. The models are trained on the **Wall Street Journal**, **LibriSpeech** or **Tedlium** and evaluated on different evaluation sets. The results on in-domain evaluation sets are in gray color. All the results are without a language model.

low-resource languages (§4.1) and compare the relative reduction in WER by switching from standard spectrogram features and the learned representations. As these are very small datasets, we trained the same **DeepSpeech2 small** architecture with the Adam optimizer with a fixed learning rate of 0.0002 and gradient clipping with maximum norm 25.0 for all languages.

Figure 2 summarizes results. Again, we find that the CPC-8k representations outperform other features by a large margin and that the models trained on the representations trained on using the audio of (English-only) LibriSpeech do not perform even as well as basic spectrogram features. This suggests that the representations learned on large-scale data capture a phonetic space that generalizes across different languages, but that diversity of linguistic inputs is crucial for developing this universality.

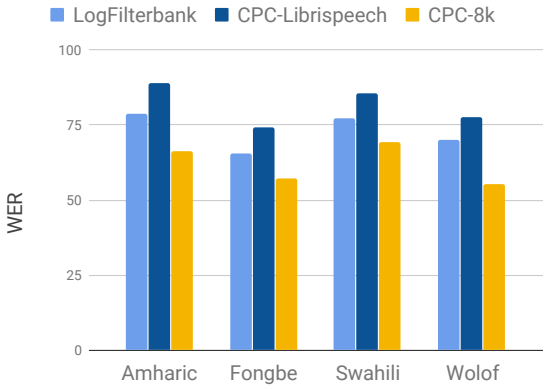

Figure 2: Speech recognition performance on low-resource African languages (in word error rate). CPC features trained on diverse datasets features significantly outperform baseline log-filterbank features whereas the features trained only on English underperform the baseline.

4.5 Multilingual Transfer

As a final exploration of the transferability of the representations, we evaluate the representations on a diverse language set of languages with varying amounts of training data and compare the relative reductions in word error rate we obtain when using standard features and switching to the CPC-8k representations. As most of the dataset are small, we trained **DeepSpeech2 small** models with the Adam optimizer with a fixed learning rate of 0.0002 and applied gradient clipping with maximum norm 25.0, using the same configuration for all languages.

Figure 3 summarizes results. Since the experiments above showed that CPC-LibriSpeech features performed badly, we only compare the relative error rediction with CPC-8k features over spectrogram features. In all cases, we find that the CPC-8k representations improve performance relative to spectorgram feature baselines. The largest improvement was obtained on Sundanese where the WER with spectrogram was 27.85 but dropped to 11.49 using CPC-8k features.

Discussion As our pre-training data did not have any language labels, it is unclear how many samples were seen for each language during pre-training. However, it is important to know that the *uncurated* multilingual pre-training can improve speech recognition performance on many languages. These results suggests, in practice, that one could use a universal speech feature extractor for many languages instead of training one for each language individually (Kannan et al., 2019).

4.6 Control: English Speech Recognition

Thus far, we have focused on robustness and transferability and seen that CPC-8k features offer considerable benefits in these dimensions compared to traditional features. It remains to demonstrate how well they work in powerful architectures where large amounts of labeled training data is available. To test this, we used 10% and 100% portions of LibriSpeech dataset to train speech recognition models, again comparing different features. Our architecture is a standard **TDNN**. The speech recognition models are trained in the similar way as standard models (Collobert et al., 2016; Kuchaiev et al., 2018). The models are trained with Adam optimizer with learning rate 0.0002 and gradient clipping with a maximum norm 5.0 together with the polynomial learning rate decay method with power 2.0 is used over 200 epochs.[5]

Table 3 summarizes the results with **TDNN** models trained on different sizes of LibriSpeech dataset. We see that even if the speech recognition models have a large number of parameters and are trained on plenty of supervised data, the learned representations still provide significant improvements. The pattern continues to hold if we use beam search decoding with a language model.[6] Our **+ LM decoding** results are comparable to the OpenSeq2Seq benchmark, since we used the exact same LM and decoding algorithm as they used (Kuchaiev et al., 2018).

Although better results contain be obtained using newer architectures than TDNN (Park et al., 2019; Synnaeve et al., 2019), it still represents a standard and important recognition architecture and the results prove that the representations learned from diverse and noisy data can improve large speech recognition model on English in both low-data and high-data regimes.

[5]These hyperparameters were chosen to give optimal performance with baseline log filterbank features, and used, unchanged for our learned features.

[6]http://www.openslr.org/resources/11/4-gram.arpa.gz

5 Related Work

Unsupervised learning played an import role in the reintroduction of deep networks to speech processing (Hinton et al., 2012), as well as other application areas (Hinton et al., 2006; Bengio et al., 2007; Vincent et al., 2010). After a period of focusing on supervised techniques, unsupervised representation learning has recently seen a resurgence in a variety of modalities (Doersch and Zisserman, 2017; van den Oord et al., 2018; Donahue and Simonyan, 2019; Bachman et al., 2019) and has led to improved results, especially in low-data regimes (Hénaff et al., 2019; Schneider et al., 2019). In natural language processing, pretrained representations can outperform state-of-the-art system even in high data regimes (Mikolov et al., 2013; Devlin et al., 2019).

The last two years have produced a large amount of work on unsupervised speech representation learning. Some of this work has been evaluated in terms of its ability to perform phone recognition and similar audio classification tasks (van den Oord et al., 2018). Like us, Schneider et al. (2019); Baevski et al. (2019) applied learned representations to speech recognition tasks and evaluated on how well in-domain WER was improved. However, as we argued in the paper, such an evaluation misses the opportunity to assess whether these systems become more robust to domain shift and to what extent the learned representations appropriate for different languages.

Finally, the ZeroSpeech challenges have explicitly looked at correlations between learned representations and phonetic structures that generalize across many languages and adapt to new speakers (Dunbar et al., 2017, 2019). Kahn et al. (2019b); Rivière et al. (2020) learned representations with contrastive predictive coding on 60,000 hours of English speech and could show that their representations are correlated well with phonetic structure of English and other languages; however, they did not evaluate these representations in a supervised speech recognizer.

Recently, there have been considerable improvements in purely supervised speech recognition systems. Data augmentation (Park et al., 2019), self-training (Synnaeve et al., 2019; Kahn et al., 2019a) have advanced the state-of-the-art performance on

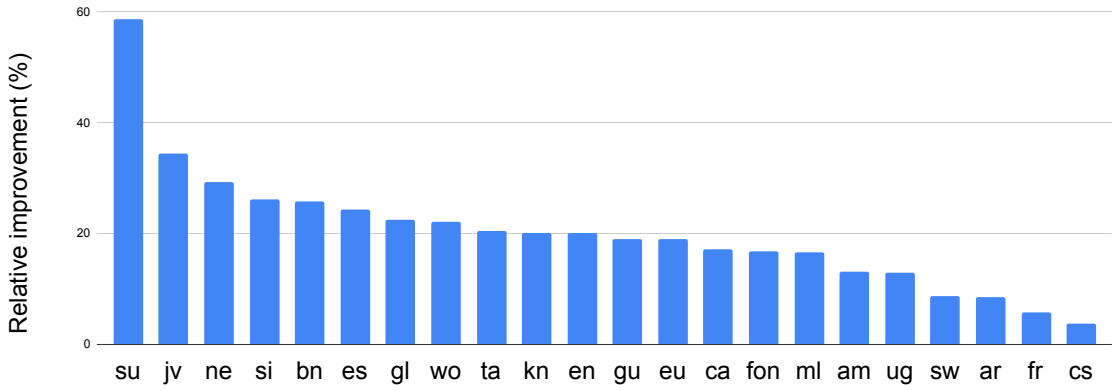

Figure 3: Relative improvements (in percentage) on speech recognition on many languages with CPC-8k features over Spectrogram features. Each column correspond to language code explained in Table 4. Note that **en** is Nigerian English and **fr** is African French.

	LibriSpeech							
	dev-clean		dev-other		test-clean		test-other	
	10%	100%	10%	100%	10%	100%	10%	100%
LibriSpeech								
LogFilterbank (OpenSeq2Seq)	-	6.67	-	18.67	-	6.58	-	19.61
LogFilterbank (ours)	19.83	6.63	38.97	18.77	19.65	6.43	41.26	20.16
CPC-LibriSpeech	15.07	6.70	33.55	19.77	14.96	6.91	36.05	21.60
CPC-8k	**13.92**	**6.20**	**30.85**	**17.93**	**13.69**	**6.25**	**32.81**	**19.10**
+ LM decoding								
LogFilterbank (OpenSeq2Seq)	-	4.75	-	13.87	-	4.94	-	15.06
LogFilterbank (ours)	12.49	4.87	28.71	14.14	12.29	5.04	31.03	15.25
CPC-LibriSpeech	9.66	4.87	24.72	14.34	9.41	5.05	26.77	16.06
CPC-8k	**8.86**	**4.35**	**22.10**	**12.96**	**8.70**	**4.72**	**24.15**	**14.47**

Table 3: Sample efficiency experiments with the **TDNN** trained and evaluated on **LibriSpeech**. The results are word error rate on the LibriSpeech development and evaluation sets. 10% vs. 100% indicates the amount of training data used. The section in **+ LM decoding** contain results with beamsearch decoding with a 4-gram language model. The underlined (OpenSeq2Seq) scores are taken from public benchmarks.[7]

English speech recognition. It is likely that augmentation methods are orthogonal to the proposed improvements on universal speech representation learning, and that one could combine both to improve results even further. Additionally, the impact of data augmentation and self-training can be further assessed in terms of its impact on robustness using the methods proposed in this paper.

6 Conclusion

We have introduced an unsupervised speech representation learning method that discovers acoustic representations from up to 8000 hours of diverse and noisy speech data. We have shown, for the first time, that such pretrained representations lead speech recognition systems to be robust to domain shifts compared to standard acoustic representations, and compared to representations trained on smaller and more domain-narrow pretraining datasets. These representations were evaluated on a standard speech recognition setup where the models are trained and evaluated on in-domain data and also on transfer tasks where the models are evaluated on out-of-domain data. We obtained consistent improvements on 25 phonetically diverse languages including tonal and low-resource languages. This suggests we are making progress toward models that implicitly discover phonetic structure from large-scale unlabelled audio signals.

References

Fréjus A. A Laleye, Laurent Besacier, Eugène C. Ezin, and Cina Motamed. 2016. First automatic fongbe continuous speech recognition system: Development of acoustic models and language models. In *Proc. FedCSIS*.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *Proc. ICML*.

Philip Bachman, R Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. In *Proc. NeurIPS*.

Alexei Baevski, Steffen Schneider, and Michael Auli. 2019. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. 2018. Mine: mutual information neural estimation. In *Proc. ICML*.

Anthony J Bell and Terrence J Sejnowski. 1995. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159.

Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. 2007. Greedy layer-wise training of deep networks. In *Proc. NeurIPS*.

Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve. 2016. Wav2letter: an end-to-end convnet-based speech recognition system. *arXiv preprint arXiv:1609.03193*.

Amita Dev and Poonam Bansal. 2010. Robust features for noisy speech recognition using mfcc computation from magnitude spectrum of higher order autocorrelation coefficients. *International Journal of Computer Applications*, 10(8):36–38.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*.

Carl Doersch and Andrew Zisserman. 2017. Multi-task self-supervised visual learning. In *Proc. ICCV*, pages 2051–2060.

Jeff Donahue and Karen Simonyan. 2019. Large scale adversarial representation learning. *arXiv preprint arXiv:1907.02544*.

Ewan Dunbar, Robin Algayres, Julien Karadayi, Mathieu Bernard, Juan Benjumea, Xuan-Nga Cao, Lucie Miskic, Charlotte Dugrain, Lucas Ondel, Alan W Black, et al. 2019. The zero resource speech challenge 2019: Tts without t. In *Proc. INTERSPEECH*.

Ewan Dunbar, Xuan Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. 2017. The zero resource speech challenge 2017. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 323–330. IEEE.

A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. 2018. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*.

Robert M Fano and David Hawkins. 1961. Transmission of information: A statistical theory of communications. *American Journal of Physics*, 29:793–794.

John S Garofolo. 1993. Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium*.

Elodie Gauthier, Laurent Besacier, Sylvie Voisin, Michael Melese, and Uriel Pascal Elingui. 2016. Collecting resources in sub-saharan african languages for automatic speech recognition: a case study of wolof. *LREC*.

Hadrien Gelas, Laurent Besacier, and Francois Pellegrino. 2012. Developments of swahili resources for an automatic speech recognition system. In *Workshop Proc. SLTU*.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. ICASSP*.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proc. ICASSP*.

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

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proc. AISTATS*.

Olivier J Hénaff, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. 2019. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018. Ted-lium 3: twice as much data and corpus repartition for experiments on speaker adaptation. In *Proc. SPECOM*.

Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, et al. 2012. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, 29.

Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proc. CVPR*.

Jacob Kahn, Ann Lee, and Awni Hannun. 2019a. Self-training for end-to-end speech recognition. *arXiv preprint arXiv:1909.09116*.

Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. 2019b. Libri-light: A benchmark for asr with limited or no supervision. *arXiv preprint arXiv:1912.07875*.

Anjuli Kannan, Arindrima Datta, Tara N Sainath, Eugene Weinstein, Bhuvana Ramabhadran, Yonghui Wu, Ankur Bapna, Zhifeng Chen, and Seungji Lee. 2019. Large-scale multilingual speech recognition with a streaming end-to-end model. In *Proc. INTERSPEECH*.

Oleksii Kuchaiev, Boris Ginsburg, Igor Gitman, Vitaly Lavrukhin, Jason Li, Huyen Nguyen, Carl Case, and Paulius Micikevicius. 2018. Mixed-precision training for nlp and speech recognition with openseq2seq. *arXiv preprint arXiv:1805.10387*.

Kshitiz Kumar, Chanwoo Kim, and Richard M Stern. 2011. Delta-spectral cepstral coefficients for robust speech recognition. In *Proc. ICASSP*.

Ralph Linsker. 1988. An application of the principle of maximum information preservation to linear systems. In *Proc. NeurIPS*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. NeurIPS*.

Joseph L Mundy, Andrew Zisserman, et al. 1992. *Geometric invariance in computer vision*, volume 92. MIT press Cambridge, MA.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *Proc. ICASSP*.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Proc. INTERSPEECH*.

Douglas B Paul and Janet M Baker. 1992. The design for the wall street journal-based csr corpus. In *Proc. ACL*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A Alemi, and George Tucker. 2019. On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*.

Morgane Rivière, Armand Joulin, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux. 2020. Unsupervised pretraining transfers well across languages. In *Proc. ICASSP*.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *Proc. INTERSPEECH*.

Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Edouard Grave, Tatiana Likhomanenko, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert. 2019. End-to-end asr: from supervised to semi-supervised learning with modern architectures. *arXiv preprint arXiv:1911.08460*.

Martha Tachbelie, Solomon Teferra Abate, and Laurent Besacier. 2014. Using different acoustic, lexical and language modeling units for asr of an under-resourced language - amharic. *Speech Communication*, 56.

Joost Van De Weijer, Theo Gevers, and Arnold WM Smeulders. 2005. Robust photometric invariant features from the color tensor. *IEEE Transactions on Image Processing*, 15(1):118–127.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(Dec):3371–3408.

Steven H Weinberger and Stephen Kunath. 2009. Towards a typology of english accents. *AACL Abstract Book*, 104.

Laurenz Wiskott and Terrence J Sejnowski. 2002. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770.

Xiaojia Zhao and DeLiang Wang. 2013. Analyzing noise robustness of mfcc and gfcc features in speaker identification. In *Proc. ICASSP*.

A Multilingual evaluation datasets

For the multilingual evaluation, we only include (labeled) datasets from OpenSLR that containing more than 1GB of audio. When there is more than one dataset available for one language, we used the largest dataset. Table 4 summarizes the multilingual dataset statistics used in our evaluation.

Language name	Code	Dataset	Hours
Amharic	am	ALFFA	18.3
Fongbe	fon	ALFFA	5.2
Swahilli	sw	ALFFA	8.9
Wolof	wo	ALFFA	16.8
Czech	cs	OpenSLR-6	15.0
Uyghur	ug	OpenSLR-22	20.2
Javanese	jv	OpenSLR-35	236.8
Sundanese	su	OpenSLR-36	265.9
Tunisian Arabic	ar	OpenSLR-46	4.5
Sinhala	si	OpenSLR-52	179.6
Bengali	bn	OpenSLR-53	172.3
Nepali	ne	OpenSLR-54	123.6
African French	fr	OpenSLR-57	13.7
Catalan	ca	OpenSLR-59	71.9
Malayalam	ml	OpenSLR-63	4.4
Tamil	ta	OpenSLR-65	5.7
Spanish	es	OpenSLR-67	19.6
Nigerian English	en	OpenSLR-70	39.5
Chilean Spanish	es	OpenSLR-71	5.7
Columbian Spanish	es	OpenSLR-72	6.1
Peruvian Spanish	es	OpenSLR-73	7.3
Basque	eu	OpenSLR-76	11.0
Galician	gl	OpenSLR-77	8.2
Gujarati	gu	OpenSLR-78	6.3
Kannada	kn	OpenSLR-79	6.7

Table 4: Summary of Multilingual Datasets.