

MAGIC DUST FOR CROSS-LINGUAL ADAPTATION OF MONOLINGUAL WAV2VEC-2.0

Sameer Khurana, Antoine Laurent, James Glass

▶ **To cite this version:**

Sameer Khurana, Antoine Laurent, James Glass. MAGIC DUST FOR CROSS-LINGUAL ADAPTATION OF MONOLINGUAL WAV2VEC-2.0. ICASSP 2022, May 2022, Singapour, Singapore. hal-03544515

HAL Id: hal-03544515
https://hal.science/hal-03544515

Submitted on 26 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

MAGIC DUST FOR CROSS-LINGUAL ADAPTATION OF MONOLINGUAL WAV2VEC-2.0

Sameer Khurana[1], *Antoine Laurent*[2], *James Glass*[1]

[1]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA
[2]LIUM - Le Mans University, France

arXiv:2110.03560v1 [cs.CL] 7 Oct 2021

ABSTRACT

We propose a simple and effective cross-lingual transfer learning method to adapt monolingual wav2vec-2.0 models for Automatic Speech Recognition (ASR) in resource-scarce languages. We show that a monolingual wav2vec-2.0 is a good few-shot ASR learner in several languages. We improve its performance further via several iterations of Dropout Uncertainty-Driven Self-Training (DUST) by using a moderate-sized unlabeled speech dataset in the target language. A key finding of this work is that the adapted monolingual wav2vec-2.0 achieves similar performance as the topline multilingual XLSR model, which is trained on fifty-three languages, on the target language ASR task.

***Index Terms*—** Cross-lingual transfer learning, self training, self-supervised Learning, ASR, adaptation

1. INTRODUCTION

Few-shot learning, the ability to train a machine to exhibit intelligent behavior via a small amount of supervision has been a long-standing research goal in Artificial Intelligence. To build few-shot learners we turn to a class of transfer learning (TL) methods that extract knowledge from vast quantities of unlabeled data to make the task of learning from a few labeled examples easier. Recently, Self-Supervised Learning (SSL) has emerged as a promising TL approach of learning from unlabeled data [1–3].

SSL [4, 5] refers to the process of Pre-Training (PT) a model on unlabeled data using an SSL task, such as masked self-prediction [2]. The Pre-Trained model is then Fine-Tuned (FT) on the target task via a few labeled examples. Hence, SSL forms the first stage of the PT then FT (PT $\rightarrow$ FT) sequential TL framework [6]. Recently, speech neural net encoders Pre-Trained using the wav2vec2 SSL framework have proven to be excellent few-shot learners for automatic speech recognition (ASR) across multiple languages [7,8]. However, wav2vec2 assumes access to massive amounts of unlabeled data for PT, which diminishes their usefulness to resource-scarce languages, where the *massive unlabeled data* assumption does not hold.

To remedy the above issue, [8] proposes xlsr, a cross-lingual sequential TL framework of the form mPT $\rightarrow$ FT, i.e., Multilingual Pre-Training of wav2vec2 followed by target language ASR fine-tuning on a few labeled examples. Indeed, Pre-Trained xlsr is an excellent few-shot learner for ASR in multiple languages. However, in this work we show that xlsr's ASR performance is quite poor if there is a domain mismatch between the target language speech and the speech data used to Pre-Train xlsr. Thus, to make xlsr a truly universal speech model, we would have to Pre-Train on speech from all languages in all possible speech domains, which is clearly an unscalable strategy. Instead, in this work, we propose a TL framework that could efficiently adapt any Pre-Trained wav2vec2 model, monolingual or multilingual, to make it a good few-shot ASR learner in any target language in any speech domain.

In this work, motivated by the SSL framework's limitations when developing ASR for a resource-scarce language, we propose a simple yet effective cross-lingual TL framework (§2) for wav2vec2 model adaptation to a target language. Our adaptation framework is a sequential TL framework consisting of three steps: First, we Pre-Train a wav2vec2 model on a high-resource language. Second, we perform supervised fine-tuning of the Pre-Trained wav2vec2 model on the target language ASR task using ten hours of labeled data. Finally, we perform Dropout Uncertainty-Driven Self-Training (DUST) [9] using a hundred hours of unlabeled speech data in the target language for adaptation of the Fine-Tuned wav2vec2 model.

Through this work, we make the following **contributions**: 1) We analyze the cross lingual transferability of several Pre-Trained English wav2vec2 models (Table 1) across eight target languages. We show that by simply fine-tuning English wav2vec2 on ten hours of labeled data in target languages, we can recover on average up to 86% of the performance of the fine-tuned multilingual xlsr topline. Still, there is a considerable gap in performance between wav2vec2 and xlsr on target languages that are considered in-domain for xlsr, but the gap is much smaller on a more challenging out-of-domain Arabic target language. Another interesting finding is that ASR Fine-Tuning of the Pre-Trained wav2vec2 models on labeled data in the source language (English) before Fine-Tuning on the target languages hurts cross-lingual transfer. 2) We adapt an English wav2vec2 model to two target languages, French and Arabic, under the constraint that in each target language we have ten hours of labeled data for ASR training and a hundred hours of unlabeled data for adaptation. For French, we show that by starting with a Pre-Trained English wav2vec2 model and applying the proposed adaptation procedure (§2), we are able to reach similar ASR performance as the xlsr topline. For Arabic, both the xlsr and English wav2vec2 perform poorly and hence, we apply the adaptation procedure to both the models and improve the ASR performance considerably. A key finding of this study is that it is possible to adapt a monolingual wav2vec2 model Pre-Trained on a high-resource language using moderately-sized unlabeled data and small-sized labeled data in the target language to achieve similar performance as the multilingual wav2vec2 model Pre-Trained on multiple languages. Although the amount of unlabeled data that we use for adaptation is orders of magnitude smaller than the data used to Pre-Train wav2vec2 models, a moderate-sized unlabeled dataset might not be available for extremely resource-scarce and endangered languages. This scenario is out of scope for this paper.

This work uses HPC resources of IDRIS under the allocation AD011012527 made by GENCI.

2. METHOD

Self-Training Self-Training (ST) [10] is a Teacher/Student (T/S) `TL` framework that leverages unlabeled data by pseudo-labeling it. ST proceeds by building a base model, known as teacher, using the labeled data. The teacher is used to predict (pseudo-)labels for the unlabeled data points. Then, a new model, known as student, is trained on the combined labeled and pseudo-labeled data points. Due to having access to more supervision, the student is expected to generalize better than the teacher on the task at hand. ST is an iterative process, where, the student from a previous round becomes the teacher for the next round of ST. Recently, ST has shown excellent results in neural sequence generation tasks such as ASR [9, 11, 12] and Machine Translation [13]

Transfer Learning Algorithm The overall transfer learning process is described in Algorithm 1. We assume access to a set $\mathcal{L}_T$ of labeled examples and a set $\mathcal{U}_T$ of unlabeled speech utterances in the target language. Also, we are given a set $\mathcal{U}_S$ of unlabeled speech utterances in the source language. The transfer learning process proceeds by Pre-training a neural network $f_{\phi,p}$ on unlabeled source language set $\mathcal{U}_S$ with dropout layers, using a dropout probability $p \in [0, 1]$. The Pre-training process leads to the initial model $f_{\phi_0,p}$, which is Fine-Tuned on the target language labeled set $\mathcal{L}_T$ to give the first-generation teacher model $f_{\phi_1,p}$ for Dropout-Uncertainty driven Self-Training (`DUST`). Next, the base teacher model $f_{\phi_1,p}$ is used to provide predictions on the target language unlabeled set $\mathcal{U}_T$ to provide pseudo-parallel data of which a subset $\mathcal{P}$ is chosen based on the model's uncertainty about its predictions on each unlabeled data point $x_u \in \mathcal{U}_T$. Finally, a student model, is trained on the combined labeled $\mathcal{L}_T$ and pseudo-labeled set $\mathcal{P}$. We perform N iterations of the Teacher/Student training, where the student $f_{\phi_n,p}$ from the n^{th} iteration becomes teacher for the $(n+1)^{th}$ iteration. Usually, in each iteration of `DUST`, a randomly initialized neural network is used as the student model, but, in our adaptation framework, the Pre-Trained source language SSL model $f_{\phi_0,p}$ is used as the student in each `DUST` iteration.

DUST performs pseudo-label filtering by measuring the model's confidence about its predictions on the unlabeled points $x_u \in \mathcal{U}_T$. The filtering process for a particular unlabeled example x_u consists of the following steps: 1) First, we generate a reference hypothesis $\hat{y}^u_{\text{ref}}$ for the unlabeled instance x_u using beam search. During inference, the model's dropout layers are deactivated and hence, this step imitates the usual ASR inference process. 2) Second, we sample R hypotheses $(\hat{y}^r_u)_{r=1}^R$ from the model by running beam search R times with a different random seed $r \in R$ each time. During inference, the dropout layers are active, hence each beam search iteration would lead to a slightly different hypothesis. This is akin to getting predictions from different models. 3) Finally, we compute the Levenshtein edit distance [14] normalized by the length of the reference hypothesis between each of the R stochastically sampled hypothesis and the one reference hypothesis, which gives us a set $\mathcal{E}$ of R edit distances. If all the edit-distances in $\mathcal{E}$ are less than the threshold ratio τ of the length $|\hat{y}^{\text{ref}}_t|$ of the reference hypothesis, then we add the pseudo-labeled data points $\{(x_u, \hat{y}^{\text{ref}}_u), (x_u, y^0_u), \ldots, (x_u, y^R_u\})$ to $\mathcal{P}$, otherwise we reject it. In practice, we set $R = 3$ and hence, we have a total of four hypotheses per utterance. Unlike the original DUST that adds only the reference pseudo-label hypothesis for x_u to the set $\mathcal{P}$, we also add the sampled hypotheses. Adding multiple pseudo-labels corresponding to an unlabeled instance x_u for student model training could increase model's robustness to noise in pseudo-labels. This idea is also explored in [15].

Algorithm 1 Transfer Learning Algorithm

Given labeled data $\mathcal{L}_S$ and unlabeled data $\mathcal{U}_S$ in the source language
Given labeled data $\mathcal{L}_T$ and unlabeled data $\mathcal{U}_T$ in the target language
Given R natural numbers
Pre-Train $f_{(\phi,p)}$ on $\mathcal{U}_S$ to get $f_{(\phi_0,p)}$
Fine-Tune $f_{(\phi_0,p)}$ on $\mathcal{L}_T$ to get $f_{(\phi_1,p)}$
for n =1 to N **do**
    $f_{(\phi_{n+1},p)}$ = `DUST`($f_{(\phi_n,p)}$, $f_{(\phi_0,p)}$, $\mathcal{L}_T$, $\mathcal{U}_T$)
end for
function `DUST`($g^{\text{Teacher}}_{(\theta,p)}$, $f^{\text{Student}}_{(\psi,p)}$, $\mathcal{L}$, $\mathcal{U}$)
    Let $\mathcal{P}$ be the set of selected pseudo-labeled data points
    Let $\mathcal{E}$ be a set of edit distances
    Initialize $\mathcal{P}$ and $\mathcal{E}$ as empty sets
    for all $x_u \in \mathcal{U}$ **do**
        Compute deterministic forward pass $g^{\text{Teacher}}_{(\theta,0)}(x_u)$
        $\hat{y}^{\text{ref}}_u = \text{beam_search}(g^{\text{Teacher}}_{(\theta,0)}(x_u))$
        for all $r \in R$ **do**
            Set random seed to r
            Compute stochastic forward pass $g^{\text{Teacher}}_{(\theta,p)}(x_u)$
            $\hat{y}^r_u = \text{beam_search}(g^{\text{Teacher}}_{(\theta,p)}(x_u))$
            $e = \text{edit_distance}(\hat{y}^r_u, \hat{y}^{\text{ref}}_u)$
            Add e to the set $\mathcal{E}$
        end for
        if $\max(\mathcal{E}) < \tau|\hat{y}^{\text{ref}}_u|$ (with τ a filtering threshold) **then**
            Add $\{(x_u, \hat{y}^{\text{ref}}_u), (x_u, \hat{y}^0_u), \ldots, (x_u, \hat{y}^R_u\})$ to $\mathcal{P}$
        end if
    end for
    Fine-Tune $f^{\text{Student}}_{(\psi,p)}$ on $\mathcal{A} = \mathcal{L} \cup \mathcal{P}$
    return $f^{\text{Student}}_{(\psi,p)}$
end function

Pre-Training In our work, we explore the following Pre-Trained `wav2vec2` SSL models that provide the initial model $f_{\phi_0,p}$ (Algorithm 1) for transfer learning.

- **Wav2Vec2.0 Base** (`w2v_base`) [7]: consists of 0.1 billion parameters and is Pre-Trained on the Librispeech 960 hours (LS960) [16] English speech dataset in the read speech domain.
- **Wav2Vec2.0 Large** (`w2v_large`) [7]: consists of 0.3 billion parameters and is Pre-Trained on either LS960 or Libri-Light 60k (LL60k) hours [17] English read speech dataset.
- **Wav2Vec2.0 Robust** (`w2v_rob`) [18]: consists of the same architecture as the large model but, is trained on three speech datasets namely Switchboard (SWBD), English part of CommonVoice (CV-En) and LL60k. We refer to the combination of these three datasets as LL60k+.
- **XLSR-53** (`xlsr`) [8]: consists of the same architecture as w2v-large which is trained on the following datasets Multilingual Speech (MLS), BABEL and CommonVoice (CV), that combined consists of 53 languages. We refer to the combination of these three datasets as MLS+.

We use the publicly available Pre-Trained `wav2vec2` model checkpoints from fairseq toolkit [19].

Table 1: Cross-Lingual Transferability of Pre-Trained `wav2vec2` model on eight target languages. Seven languages are from the MLS dataset of read audiobooks, while Arabic is from the MGB broadcast news dataset

		WER / CER [%]										WERR / CERR
Target Langs		MLS/en	MLS/fr	MLS/de	MLS/it	MLS/pl	MLS/es	MLS/pt	MLS/nl	MGB/ar		
Model	PT										Avg.↓	Avg.↑
Baseline		119.1 / 58.5	114.2 / 51.6	106.0 / 41.5	99.5 / 35.0	111.9 / 44.7	99.5 / 37.3	107.0 / 45.3	108.8 / 50.0	112.0 / 51.5	107.4 / 44.6	0 / 0
`w2v_base`	LS960	23.4 / 8.1	44.0 / 14.5	28.6 / 6.9	34.1 / 7.3	35.6 / 6.9	37.2 / 8.6	41.1 / 10.9	47.2 / 14.2	47.4 / 15.1	39.4 / 10.6	79.0 / 87.8
`w2v_large`	LS960	17.1 / 5.8	40.9 / 13.3	28.3 / 6.8	33.3 / 6.9	32.0 / 6.2	23.6 / 5.6	38.6 / 10.2	45.0 / 13.3	42.7 / 14.2	35.6 / 9.6	83.6 / 90.5
`w2v_large`	LL60k	12.3 / 4.0	39.9 / 12.7	26.7 / 6.4	31.8 / 6.7	32.8 / 6.4	21.9 / 5.1	35.6 / 9.4	42.6 / 12.6	42.0 / 13.2	34.2 / 9.1	85.6 / 92.1
`w2v_rob`	LL60k+	12.8 / 4.2	38.3 / 12.3	26.7 / 6.4	30.3 / 6.2	34.2 / 6.6	22.9 / 5.3	34.2 / 8.9	39.1 / 11.8	41.6 / 13.1	33.4 / 8.8	**86.3 / 92.5**
`w2v_large_sup`	LL60k	**7.6 / 2.5**	44.2 / 14.2	31.1 / 7.2	37.7 / 7.8	46.5 / 9.0	28.1 / 6.4	40.8 / 10.4	51.3 / 15.3	50.6 / 15.8	41.3 / 10.8	78.8 / 88.6
Topline (`xlsr`)	MLS+	17.6 / 6.3	**19.7 / 6.5**	**11.1 / 3.1**	**17.1 / 3.6**	**16.4 / 3.3**	**7.9 / 2.1**	**20.4 / 5.3**	**21.7 / 6.3**	**37.9 / 12.0**	**19.0 / 5.3**	100 / 100

Fine-Tuning The Fine-Tuning of Pre-Trained SSL models consists of 1) Adding a linear projection layer $h_\alpha : \mathbb{R}^{T\times d} \rightarrow \mathbb{R}^{T\times|V|}$ to the output of the SSL model, where V is the output character vocabulary for the task of ASR, 2) ASR task Fine-Tuning of only the projection layer for the first k training iterations and 3) Joint ASR task Fine-Tuning of both the SSL model and the projection layer until convergence. Note the `wav2vec2` SSL models consists of a Convolutional Neural Network (CNN) feature extractor, followed by a transformer encoder. The CNN feature extractor remains frozen throughout the ASR Fine-Tuning process.

3. EXPERIMENT SETUP

Transfer Learning Targets We chose seven languages from the MLS dataset as the targets for cross-lingual adaptation of the Pre-Trained `wav2vec2` SSL models, namely French (MLS/fr), German (MLS/de), Italian (MLS/it), Polish (MLS/pl), Spanish (MLS/es), Portugese (MLS/pt) and Dutch (MLS/nl). In addition, we also target Arabic from the Multi-Genre Multi-Dialectal Broadcast News (MGB) dataset [20]. In order to simulate the resource-scarce ASR scenario, we assume access to just ten hours of labeled data and a hundred hours of unlabeled data in each target language. We use the official nine hours labeled split in MLS for training and the one hour split for validation. We report Word Error Rates (WERs) on the unseen development set. The hundred hours unlabeled set is sampled randomly from the full training set (minus the utterances in the ten hours split). For Arabic, we randomly sample ten hours of labeled data, of which nine hours is used for training and one hour for validation. We also randomly sample a hundred hours of speech from the 1200 hours MGB training set for cross-lingual adaptation. The results are reported on the standard development set. For the `xlsr` model, MGB/ar is considered an out-of-domain target language because `xlsr` is Pre-Trained on multiple datasets including MLS, which are in the read speech and conversational domains, while MGB is in the broadcast news domain. This is evident from the high WERs of the Fine-Tuned `xlsr` on the MGB/ar dataset as compared to the MLS target languages in Table 1.

Hyperparameters For ASR Fine-Tuning ASR Fine-Tuning of the Pre-Trained SSL models is performed on the ten hours labeled data $(x, y) \in \mathcal{L}_T$ in the target language T, where x is the input speech waveform and y is the corresponding sub-word token sequence. We choose characters as sub-word units for ASR training. The model is trained using the Connectionist Temporal Classification (CTC) [21] loss. For optimization, we use the Adam optimizer with a learning rate schedule given by the following equation:

$$\text{lr} = \text{max_lr} * \text{warmup_steps}^{0.5} * \min(\text{step}^{-0.5}, \text{step} * \text{warmup_steps}^{-1.5})$$

where, max_lr is the maximum learning rate, warmup_steps are the number of training iterations before the maximum learning rate is achieved and step refers to the current training iteration. We use a relatively small value of 1e-4 for max_lr and the first 8k training iterations for warmup. The model is trained for a total of 300 epochs. For the first 4k training iterations, we only train the linear projection layer h_α. Batching is performed by pooling together raw speech waveforms in such a way that the total number of samples do not exceed 3.2 million. We use a gradient accumulation factor of four to ensure that the model is updated after every four training iterations, which leads to an effective batch size that is four times the original. The feature sequence output by the CNN encoder of the SSL models is randomly masked in the time dimension. We mask a span of ten consecutive time steps with a masking probability of 0.65, which leads to 65% of the input signal being masked. We use 4 V100-32GB GPUs for fine-tuning. We use the Espnet2 codebase [22] to perform all our experiments.

Table 2: Transfer of Pre-Trained `w2v_rob` to the target French language in the MLS dataset

		WER [%]		WERR [%]
Method	$\lvert\mathcal{P}\rvert$ [k]	$\mathcal{P}$	MLS / fr	MLS / fr
Baseline (`w2v_rob`)			38.3	0
DUST1	11	20.2	31.9	34.4
DUST2	24	20.3	27.4	58.6
DUST3	30	20.0	24.2	75.8
DUST4	30	19.2	23.5	79.6
DUST5	30	**18.7**	**22.3**	**86.0**
Topline (`xlsr`)			19.7	100

Decoding We use beam search decoding without a language model (LM) with a beam size of 10. We do not use an LM because, in this work we are solely concerned about the acoustic model adaptation. Also, in a resource-scarce ASR scenario, we might not have text data to train a LM.

4. RESULTS

In **Table 1**, we show the cross-lingual transferability of different Pre-Trained `wav2vec2` models on eight target languages. The goal is to

Table 3: Transfer of Pre-Trained `w2v_rob` and `xlsr` models to the target Arabic Language in the MGB dataset

Method	$\|\mathcal{P}\|$ [k]	WER [%] $\mathcal{P}$	WER [%] MGB / ar
Baseline (`w2v_rob`)			41.6
DUST1	12	21.0	32.7
DUST2	26	21.2	27.4
DUST3	30	20.8	25.2
DUST4	30	19.5	23.1
DUST5	30	**18.7**	**21.2**
`xlsr`			37.9
Baseline (`xlsr`)			37.9
DUST1	13	20.3	31.1
DUST2	29	20.4	26.3
DUST3	30	20.1	24.1
DUST4	30	18.5	22.5
DUST5	30	**18.1**	**20.8**

analyze how much of the multilingual `xlsr` topline's performance can be recovered by simply Fine-Tuning the English `wav2vec2` models on ten hours of labeled data in target languages. We Fine-Tune a randomly initialized transformer encoder which consists of the same architecture as `w2v_base` on ten hours of labeled data in each language to use as a baseline. We perform ASR Fine-Tuning of several Pre-Trained English `wav2vec2` on ten hours of labeled data in target languages and compare their ASR performance against the Fine-Tuned `xlsr` model topline. We make the following conclusions: 1) **Pre-Training Matters**: ASR Fine-Tuning of Pre-Trained English `wav2vec2` models lead to significant improvements in WERs on target languages over the baseline. Through the simple `PT` $\rightarrow$ `FT` process, we are able to recover on average 79% to 86% of the WER and 88% to 93% of the CER of the `xlsr` topline. 2) **Big SSL models provide better transfer**: By Fine-Tuning `w2v_large` that is Pre-Trained on the LS960 dataset, we are able to recover on average 83% of the topline WER compared to 79% achieved by Fine-Tuning `w2v_base` that is also Pre-Trained on LS960. 3) **Pre-Training dataset size matters upto a point**: Fine-Tuned `w2v_large` that is Pre-Trained on LL60k recovers on average 86% of the topline WER compared to 84% recovered by Fine-Tuning `w2v_large` that is Pre-Trained on LS960. But the gap in average Word Error Rate Recovery (WERR) between `w2v_rob` that is Pre-Trained on the combined CV, SWBD and LL60k datasets, and `w2v_large` that is Pre-Trained only on LL60k is less than one percentage point (pp). 4) **ASR Fine-Tuning of SSL models on source language hurts transfer**: The average WERR on target languages of `w2v_large_sup` model which is Pre-Trained on LL60k followed by its ASR Fine-Tuning on labeled LS960 is worse than directly Fine-Tuning the Pre-Trained `wav2vec2` models on the target languages. The WERR for `w2v_large_sup` is about 8pp worse than `w2v_rob` that is directly Fine-Tuned on target languages. 5) **About the out-of-domain Arabic Target Language**: We see that on the seven in-domain languages (MLS/x, where x is the target language) `xlsr` achieves an average WER of 16.5% compared to 29.8% achieved by the ASR Fine-Tuning of `w2v_rob`, the best of the English `wav2vec2` models, giving a performance gap of about 14pp between the two. However, on the out-of-domain Arabic target language (MGB/ar), the gap is less than 4pp. Next, we perform cross-lingual adaptation of Pre-Trained `wav2vec2` models using DUST. We choose French and Arabic as the target languages for transfer learning and `w2v_rob` and `xlsr` as the target models for adaptation.

In **Table 2**, we use DUST to perform cross-lingual adaptation of Pre-Trained `w2v_rob` to French (MLS/fr). DUST proceeds as follows: 1) First, we perform the ASR Fine-Tuning of the initial `w2v_rob` ($f_{\phi_0,p}$) model using the standard nine hours labeled split provided by MLS/fr dataset to get the first-generation teacher $f_{\phi_1,p}$ (§2). 2) Second, $f_{\phi_1,p}$ is used to generate pseudo-labels on the random 100 hours unlabeled split from MLS/fr, which amounts to about 30k utterances, using the pseudo-label generation process explained in §2 to give a set $\mathcal{P}$ of pseudo-parallel data. We use 0.2 as the value of the DUST filtering threshold τ. We choose τ blindly without tuning it on a labeled validation set. 3) Lastly, we Fine-Tune `w2v_rob` (student), $f_{\phi_0,p}$, on the combined labeled and pseudo-labeled data $\mathcal{P}$ to get $f_{\phi_2,p}$, which is used as the teacher for the next iteration of DUST. We perform a total of five DUST iterations. The final student model $f_{\phi_5,p}$ achieves a WER of 22.3% which is 16pp lower than the WER of 38.3% achieved by the first generation teacher model $f_{\phi_1,p}$. Furthermore, $f_{\phi_5,p}$ is able to recover 86% of the `xlsr` topline's WER. Additionally, we make the following observations: 1) Unsurprisingly, the size of the filtered pseudo-label set $\mathcal{P}$ (denoted as $|\mathcal{P}|$ in Table 2) is larger in later DUST iterations due to the continual improvement in the quality of the student (see WER [%] in Table 2), which leads to an improved teacher for subsequent DUST iterations; an improved teacher leads to cleaner pseudo-labels and hence less rejected unlabeled data points during the pseudo-label filtering process. 2) Also, in the later DUST iterations the quality of the pseudo-labels improve, which is implied by the lower WER on pseudo-label set $\mathcal{P}$ during the later iterations. Next, we consider Arabic (MGB/ar) as the target language for transfer learning, a more challenging transfer learning scenario.

In **Table 3**, we perform adaptation of `w2v_rob` and `xlsr` to the MGB/ar dataset. Here, the results are achieved by following the same adaptation process detailed above for experiments in Table 2. After five DUST iterations, we achieve the final WER of 20.8% when starting with a Fine-Tuned `xlsr` model as the first generation teacher $f_{\phi_1,p}$. This result is about 17pp better than the WER of 37.4% with $f_{\phi_1,p}$. Similar improvements are achieved when using the Fine-Tuned `w2v_rob` as $f_{\phi_1,p}$ for DUST iterations.

5. CONCLUSIONS

We conclude by summarizing the key findings of the paper. We show (Table 1) that the monolingual Pre-Trained `wav2vec2` models transfer well across multiple languages. In particular, we show that by performing ASR Fine-Tuning of `wav2vec2_robust` on ten hours of labeled data in a target language we are able to recover on average 86% of the performance of the topline multilingual `xlsr` model that is Pre-Trained on 53 languages and Fine-Tuned on the same amount of labeled target language data. This finding concurs with similar findings of [23] on cross-lingual transfer of monolingual Pre-Trained SSL models to different target languages for the task of phoneme recognition. Our work goes a step further and proposes a simple yet effective cross-lingual transfer learning algorithm (§2) for adaptation of monolingual `wav2vec2` models via Dropout Uncertainty-Driven Self-Training (DUST) by leveraging hundred hours of unlabeled speech data from the target language. We show (Table 2) that DUST improves over the baseline model that is Fine-Tuned only on labeled target language data, and is able to recover 86% of the WER of the topline `xlsr` model when adapting to French. We show similar results (Table 3) when considering Arabic as the target language. Future work should explore combining our method with the adapter framework for cross-lingual transfer learning [24–27].

6. REFERENCES

[1] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[3] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2019.

[4] V. R. DeSa, "Learning classification with unlabeled data," in *Proceedings of the 6th International Conference on Neural Information Processing Systems*, ser. NIPS'93. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993, p. 112–119.

[5] J. Schmidhuber, "Making the world differentiable: On using self-supervised fully recurrent neural networks for dynamic reinforcement learning and planning in non-stationary environments," Tech. Rep., 1990.

[6] D. Wang and T. F. Zheng, "Transfer learning for speech and language processing," 2015.

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

[8] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," 2020.

[9] S. Khurana, N. Moritz, T. Hori, and J. L. Roux, "Unsupervised domain adaptation for speech recognition via uncertainty driven self-training," *Proc. ICASSP*, 2021.

[10] H. Scudder, "Probability of error of some adaptive pattern-recognition machines," *IEEE Trans. Inf. Theory*, vol. 11, no. 3, pp. 363–371, 1965.

[11] J. Kahn, A. Lee, and A. Hannun, "Self-training for end-to-end speech recognition," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020. [Online]. Available: http://dx.doi.org/10.1109/ICASSP40776.2020.9054295

[12] Q. Xu, T. Likhomanenko, J. Kahn, A. Hannun, G. Synnaeve, and R. Collobert, "Iterative pseudo-labeling for speech recognition," *arXiv preprint arXiv:2005.09267*, 2020.

[13] J. He, J. Gu, J. Shen, and M. Ranzato, "Revisiting self-training for neural sequence generation," *arXiv preprint arXiv:1909.13788*, 2019.

[14] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, vol. 10, p. 707, Feb. 1966.

[15] S. Dey, P. Motlicek, T. Bui, and F. Dernoncourt, "Exploiting semi-supervised training through a dropout regularization in end-to-end speech recognition," *arXiv preprint arXiv:1908.05227*, 2019.

[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, Apr. 2015.

[17] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7669–7673, https://github.com/facebookresearch/libri-light.

[18] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," 2021.

[19] 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," 2019.

[20] A. Ali, P. Bell, J. Glass, Y. Messaoui, H. Mubarak, S. Renals, and Y. Zhang, "The mgb-2 challenge: Arabic multi-dialect broadcast media recognition," 2019.

[21] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:abs/1211.3711*, 2012.

[22] S. Watanabe, F. Boyer, X. Chang, P. Guo, T. Hayashi, Y. Higuchi, T. Hori, W.-C. Huang, H. Inaguma, N. Kamo, S. Karita, C. Li, J. Shi, A. S. Subramanian, and W. Zhang, "The 2020 espnet update: new features, broadened applications, performance improvements, and future plans," 2020.

[23] M. Rivière, A. Joulin, P.-E. Mazaré, and E. Dupoux, "Unsupervised pretraining transfers well across languages," 2020.

[24] J. Pfeiffer, I. Vulić, I. Gurevych, and S. Ruder, "Mad-x: An adapter-based framework for multi-task cross-lingual transfer," 2020.

[25] 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," 2019.

[26] S. Kessler, B. Thomas, and S. Karout, "Continual-wav2vec2: an application of continual learning for self-supervised automatic speech recognition," 2021.

[27] W. Hou, H. Zhu, Y. Wang, J. Wang, T. Qin, R. Xu, and T. Shinozaki, "Exploiting adapters for cross-lingual low-resource speech recognition," 2021.