Robust Speech Recognition via Large-Scale Weak Supervision

Alec Radford [*1] Jong Wook Kim [*1] Tao Xu [1] Greg Brockman [1] Christine McLeavey [1] Ilya Sutskever [1]

Abstract

We study the capabilities of speech processing systems trained simply to predict large amounts of transcripts of audio on the internet. When scaled to 680,000 hours of multilingual and multitask supervision, the resulting models generalize well to standard benchmarks and are often competitive with prior fully supervised results without the need for any dataset specific fine-tuning. When compared to humans, the models approach their accuracy and robustness. We are releasing models and inference code to serve as a foundation for further work on robust speech processing.

1. Introduction

Progress in speech recognition has been energized by the development of unsupervised pre-training techniques exemplified by Wav2Vec 2.0 (Baevski et al., 2020). Since these methods learn directly from raw audio without the need for human labels, they can productively use large datasets of unlabeled speech and have been quickly scaled up to 1,000,000 hours of training data (Zhang et al., 2021), far more than the 1,000 or so hours typical of an academic supervised dataset. When fine-tuned on standard benchmarks, this approach has improved the state of the art, especially in a low-data setting.

These pre-trained audio encoders learn high-quality representations of speech, but because they are purely unsupervised they lack an equivalently performant decoder mapping those representations to usable outputs, necessitating a fine-tuning stage in order to actually perform a task such as speech recognition[1]. This unfortunately limits their usefulness and impact as fine-tuning can still be a complex process requiring a skilled practitioner. There is an additional risk with requiring fine-tuning. Machine learning methods are exceedingly adept at finding patterns within a training dataset which boost performance on held-out data from the same dataset. However, some of these patterns are brittle and spurious and don't generalize to other datasets and distributions. In a particularly disturbing example, Radford et al. (2021) documented a 9.2% increase in object classification accuracy when fine-tuning a computer vision model on the ImageNet dataset (Russakovsky et al., 2015) without observing any improvement in average accuracy when classifying the same objects on seven other natural image datasets. A model that achieves "superhuman" performance when trained on a dataset can still make many basic errors when evaluated on another, possibly precisely because it is exploiting those dataset-specific quirks that humans are oblivious to (Geirhos et al., 2020).

This suggests that while unsupervised pre-training has improved the quality of audio encoders dramatically, the lack of an equivalently high-quality pre-trained decoder, combined with a recommended protocol of dataset-specific fine-tuning, is a crucial weakness which limits their usefulness and robustness. The goal of a speech recognition system should be to work reliably "out of the box" in a broad range of environments without requiring supervised fine-tuning of a decoder for every deployment distribution.

As demonstrated by Narayanan et al. (2018), Likhomanenko et al. (2020), and Chan et al. (2021) speech recognition systems that are pre-trained in a *supervised* fashion across many datasets/domains exhibit higher robustness and generalize much more effectively to held-out datasets than models trained on a single source. These works achieve this by combining as many existing high-quality speech recognition datasets as possible. However, there is still only a moderate amount of this data easily available. SpeechStew (Chan et al., 2021) mixes together 7 pre-existing datasets totalling 5,140 hours of supervision. While not insignificant, this is still tiny compared to the previously mentioned 1,000,000 hours of unlabeled speech data utilized in Zhang et al. (2021).

Recognizing the limiting size of existing high-quality supervised datasets, recent efforts have created larger datasets for speech recognition. By relaxing the requirement of gold-standard human-validated transcripts, Chen et al. (2021) and

[*] Equal contribution [1] OpenAI, San Francisco, CA 94110, USA. Correspondence to: Alec Radford <alec@openai.com>, Jong Wook Kim <jongwook@openai.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1] Baevski et al. (2021) is an exciting exception - having developed a fully unsupervised speech recognition system

Galvez et al. (2021) make use of sophisticated automated pipelines to scale weakly supervised speech recognition to 10,000 and 30,000 hours of noisier training data. This trade-off between quality and quantity is often the right call. Although understudied so far for speech recognition, recent work in computer vision has demonstrated that moving beyond gold-standard crowdsourced datasets such as ImageNet (Russakovsky et al., 2015) to much larger but weakly supervised datasets significantly improves the robustness and generalization of models (Mahajan et al., 2018; Kolesnikov et al., 2020).

Yet these new datasets are only a few times larger than the sum of existing high-quality datasets and still much smaller than prior unsupervised work. In this work we close that gap, scaling weakly supervised speech recognition the next order of magnitude to 680,000 hours of labeled audio data. We call our approach Whisper. We demonstrate models trained at this scale transfer well to existing datasets zero-shot, removing the need for any dataset-specific fine-tuning to achieve high-quality results.

In addition to scale, our work also focuses on broadening the scope of weakly supervised pre-training beyond English-only speech recognition to be both multilingual and multitask. Of those 680,000 hours of audio, 117,000 hours cover 96 other languages. The dataset also includes 125,000 hours of `X→en` translation data. We find that for sufficiently large models there is no drawback and even benefits to joint multilingual and multitask training.

Our work suggests that simple scaling of weakly supervised pre-training has been underappreciated so far for speech recognition. We achieve these results without the need for the self-supervision or self-training techniques that have been a mainstay of recent large-scale speech recognition work.

2. Approach

2.1. Data Processing

Relying on the expressiveness of sequence-to-sequence models to learn to map between utterances and their transcribed form, we train Whisper models to predict the raw text of transcripts without any significant standardization or pre-processing. This simplifies the speech recognition pipeline since it removes the need for a separate inverse text normalization step to output naturalistic transcriptions.

We construct the dataset from audio that is paired with transcripts on the Internet. This results in a very diverse dataset covering a broad distribution of audio from many different environments, recording setups, speakers, and languages. While diversity in audio quality can help train a model to be robust, diversity in transcript quality is not similarly beneficial. Initial inspection showed a large amount of subpar transcripts in the raw dataset. To address this, we developed several automated filtering methods to improve transcript quality.

Many transcripts on the internet are not human-generated but the output of existing ASR systems. We use various heuristics based on punctuation, capitalization, and other features to detect and remove machine-generated transcripts from the training dataset. While many ASR systems include some level of inverse text normalization, it is often still detectable from some give away such as never including commas.

We also use an audio language detector, which was created by fine-tuning a model trained on a prototype version of the dataset on VoxLingua107 (Valk & Alumäe, 2021) to ensure that the spoken language matches the language of the transcript according to CLD2. If the two do not match, we don't include the (audio, transcript) pair as a speech recognition training example in the dataset. We make an exception if the transcript language is English and add these pairs to the dataset as `X→en` speech translation training examples instead. We use fuzzy de-duping of transcript texts to reduce the amount of duplication and automatically generated content in the training dataset.

We break audio files into 30-second segments paired with the subset of the transcript that occurs within that time segment. We train on all audio, including segments where there is no speech (though with sub-sampled probability) and use these segments as training data for voice activity detection.

For an additional filtering pass, after training an initial model we aggregated information about its error rate on training data sources and performed manual inspection of these data sources sorting by a combination of both high error rate and data source size in order to identify and remove low-quality ones efficiently. This inspection showed a large amount of only partially transcribed or poorly aligned/misaligned transcripts as well as remaining low-quality machine-generated captions that filtering heuristics did not detect.

To avoid contamination, we perform de-duplication at a transcript level between the training dataset and the evaluation datasets we thought were at higher risk of overlap, namely TED-LIUM 3 (Hernandez et al., 2018).

2.2. Model

Since the focus of our work is on studying the capabilities of large-scale supervised pre-training for speech recognition, we use an off-the-shelf architecture to avoid confounding our findings with model improvements. We chose an encoder-decoder Transformer (Vaswani et al., 2017) as this architecture has been well validated to scale reliably. All

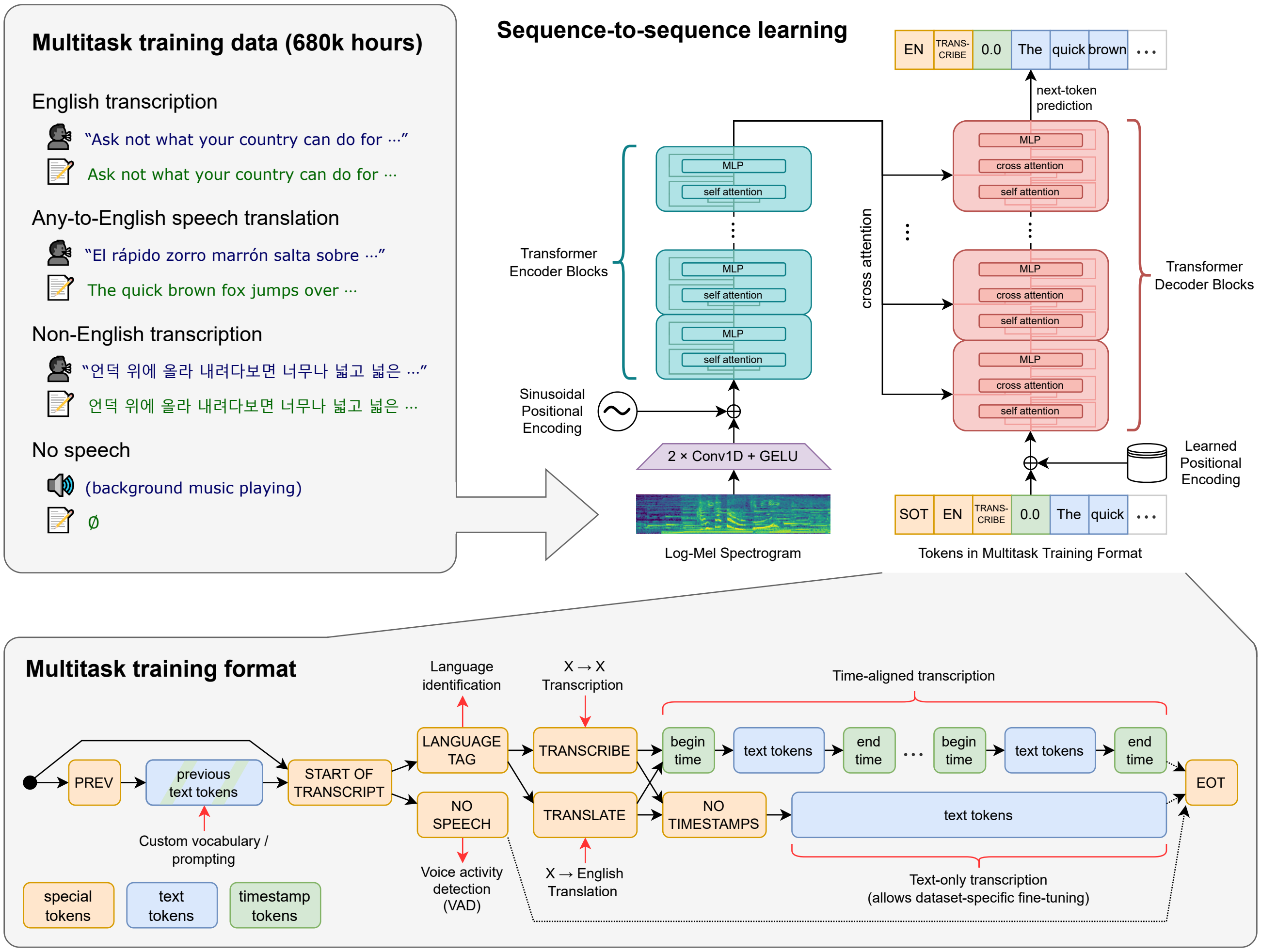
Multitask training data (680k hours)
English transcription
"Ask not what your country can do for ⋯"
Ask not what your country can do for ⋯
Any-to-English speech translation
"El rápido zorro marrón salta sobre ⋯"
The quick brown fox jumps over ⋯
Non-English transcription
"언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯"
언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯
No speech
(background music playing)
∅
Sequence-to-sequence learning
EN
TRANS-CRIBE
0.0
The
quick
brown
⋯
next-token prediction
MLP
self attention
cross attention
Transformer Encoder Blocks
Transformer Decoder Blocks
cross attention
Sinusoidal Positional Encoding
2 × Conv1D + GELU
Log-Mel Spectrogram
Learned Positional Encoding
SOT
EN
TRANS-CRIBE
0.0
The
quick
⋯
Tokens in Multitask Training Format
Multitask training format
Language identification
X → X Transcription
Time-aligned transcription
PREV
previous text tokens
START OF TRANSCRIPT
LANGUAGE TAG
NO SPEECH
TRANSCRIBE
TRANSLATE
begin time
text tokens
end time
⋯
begin time
text tokens
end time
NO TIMESTAMPS
text tokens
EOT
Custom vocabulary / prompting
Voice activity detection (VAD)
X → English Translation
Text-only transcription (allows dataset-specific fine-tuning)
special tokens
text tokens
timestamp tokens

Figure 1. **Overview of our approach.** A sequence-to-sequence Transformer model is trained on many different speech processing tasks, including multilingual speech recognition, speech translation, spoken language identification, and voice activity detection. All of these tasks are jointly represented as a sequence of tokens to be predicted by the decoder, allowing for a single model to replace many different stages of a traditional speech processing pipeline. The multitask training format uses a set of special tokens that serve as task specifiers or classification targets, as further explained in Section 2.3.

audio is re-sampled to 16,000 Hz, and an 80-channel log-magnitude Mel spectrogram representation is computed on 25-millisecond windows with a stride of 10 milliseconds. For feature normalization, we globally scale the input to be between -1 and 1 with approximately zero mean across the pre-training dataset. The encoder processes this input representation with a small stem consisting of two convolution layers with a filter width of 3 and the GELU activation function (Hendrycks & Gimpel, 2016) where the second convolution layer has a stride of two. Sinusoidal position embeddings are then added to the output of the stem after which the encoder Transformer blocks are applied. The transformer uses pre-activation residual blocks (Child et al., 2019), and a final layer normalization is applied to the encoder output. The decoder uses learned position embeddings and tied input-output token representations (Press & Wolf, 2017). The encoder and decoder have the same width and number of transformer blocks. Figure 1 summarizes the model architecture.

We use the byte-level BPE text tokenizer from GPT-2 (Sennrich et al., 2015; Radford et al., 2019) for English models and refit the vocabulary (but keep the same size) for the multilingual models to avoid token fragmentation on other languages since the GPT-2 BPE vocabulary is English only.

2.3. Multitask Format

A full speech processing system involves many components in addition to speech recognition such as voice activity detection, speaker diarization, and inverse text normalization. These components are often handled by separate pipelines, resulting in a relatively complex system of interacting parts.

For simplicity, it would be ideal to have a single model perform the entire pipeline. Since there are many different tasks that can be performed on the same input audio signal: transcription, translation, voice activity detection, alignment, and language identification, some form of task specification is necessary. We specify all tasks and conditioning information as a sequence of input tokens to the decoder. We also train it to condition on the history of text of the transcript in the hope that it will learn to use longer-range text context to resolve ambiguous audio and with some probability we add the transcript text preceding the current audio segment to the decoder's context. We indicate the beginning of prediction with a `<|startoftranscript|>` token. First, we predict the language being spoken which is represented by a token for each language in our training set (99 total). These language targets are sourced from the aforementioned VoxLingua107 model. In the case where there is no speech in an audio segment, the model is trained to predict a `<|nospeech|>` token. The next token specifies the task with a `<|transcribe|>` or `<|translate|>` token. After this, we specify whether to predict timestamps or not by including a `<|notimestamps|>` token for that case. At this point, the task and desired format is fully specified, and output tokens begin. For timestamp prediction, we predict time relative to the current audio segment, quantizing all times to the nearest 20 milliseconds which matches the time resolution of Whisper models, and add additional tokens to our vocabulary for each of these. We interleave their prediction with the caption tokens: the start time token is predicted before each caption's text, and the end time token is predicted after. When a transcript segment is only partially included in the current 30-second audio chunk, we predict only the start time token for the segment when in timestamp mode, to indicate that the subsequent decoding should be performed on an audio window aligned with that time, otherwise we truncate the audio to not include the segment. Lastly, we add a `<|endoftranscript|>` token. We only mask out the training loss over the previous context text, and train the model to predict all other tokens. Please see Figure 1 for an overview of our format and training setup.

2.4. Training Details

We train a suite of models in order to study the scaling properties of Whisper ranging from 39M to 1550M params. Models were trained with AdamW (Loshchilov & Hutter, 2017) and gradient norm clipping (Pascanu et al., 2013) with a linear learning rate decay to zero after a warmup over the first 2048 updates. A batch size of 256 segments was used, and the models are trained for 2^{20} updates which is between two and three passes over the dataset. Due to only training for a few epochs, over-fitting is not a large concern, and we do not use any data augmentation or regularization and instead rely on the diversity contained within such a large dataset to encourage generalization and robustness. Please see Appendix H for full training hyperparameters.[2]

During early development and evaluation we observed that Whisper models had a tendency to transcribe plausible but almost always incorrect guesses for the names of speakers. This is due to many training transcripts including the name of the person who is speaking. To avoid this, we fine-tune Whisper models briefly on the subset of transcripts that do not include speaker annotations which removes this behavior.

3. Experiments

3.1. Zero-shot Evaluation

The goal of Whisper is to develop a single robust speech processing system that works reliably without the need for dataset specific fine-tuning to achieve high-quality results on specific distributions. To study this capability, we re-use a wide set of existing speech datasets to check whether Whisper is able to generalize well across domains, tasks, and languages. Instead of using the standard evaluation protocol for these datasets, which include both a train and test split, we evaluate Whisper in a zero-shot setting without using any of the training data for each of these datasets in order to measure robust generalization.

3.2. Evaluation Metrics

Speech recognition research typically evaluates systems based on the word error rate (WER) metric. However, WER penalizes all differences between the model's output and the reference transcript including innocuous differences in transcript style. Systems that output transcripts that humans consider correct can still have a large WER due to minor formatting differences. This issue is particularly acute for zero-shot models like Whisper, which do not observe any examples of specific datasets transcript formats.

We address this problem with standardization of text before the WER calculation to minimize penalization of non-semantic differences. Our text normalizer was developed through iterative manual inspection to identify common patterns where naive WER penalized Whisper models for an innocuous difference. Appendix D includes full details. We caution this normalization comes at a risk of overfitting to the transcription style of Whisper models which we investigate in Appendix D.

[2] After the original release of Whisper, we trained an additional Large model (denoted V2) for 2.5X more epochs while adding SpecAugment (Park et al., 2019), Stochastic Depth (Huang et al., 2016), and BPE Dropout (Provilkov et al., 2019) for regularization. Reported results have been updated to this improved model unless otherwise specified.

3.3. English Speech Recognition

In 2015, Deep Speech 2 (Amodei et al., 2015) reported a speech recognition system matched human-level performance when transcribing the LibriSpeech test-clean split. Seven years later the SOTA WER on LibriSpeech test-clean has dropped another 73% from their 5.3% to 1.4% (Zhang et al., 2021), far below their reported human-level error rate of 5.8%. Despite this massive improvement in performance on held-out but in-distribution data, speech recognition models trained on LibriSpeech remain far above human error rates when used in other settings. What explains this gap between reportedly superhuman performance in-distribution and subhuman performance out-of-distribution?

We suspect a large part of this gap between human and machine behavior is due to conflating different capabilities being measured by human and machine performance on a test set. The difference arises not in the testing but in how they trained for it. Humans are often asked to perform a task given little to no supervision on the specific data distribution being studied. Thus human performance is a measure of out-of-distribution generalization. But machine learning models are usually evaluated after training on a large amount of supervision from the evaluation distribution, meaning that machine performance is instead a measure of in-distribution generalization. While both humans and machines are being evaluated on the same *test* data, two quite different abilities are being measured due to a difference in *train* data.

Whisper models, which are trained on a broad and diverse distribution of audio and evaluated in a zero-shot setting, could potentially match human behavior much better than existing systems. To study whether this is the case (or whether the difference between machine and human performance is due to yet-to-be-understood factors) we can compare Whisper models with both human performance and standard fine-tuned machine learning models and check which they more closely match.

To quantify this difference, we examine both *overall* robustness, that is average performance across many distributions/datasets, and *effective* robustness, introduced by Taori et al. (2020), which measures the difference in expected performance between a reference dataset, which is usually in-distribution, and one or more out-of-distribution datasets. A model with high effective robustness does better than expected on out-of-distribution datasets as a function of its performance on the reference dataset and approaches the ideal of equal performance on all datasets. For our analysis, we use LibriSpeech as the reference dataset due to its central role in modern speech recognition research and the availability of many released models trained on it, which allows for characterizing robustness behaviors. We use a suite of 12 other academic speech recognition datasets to study out-of-distribution behaviors. Full details about these datasets can be found in Appendix B.

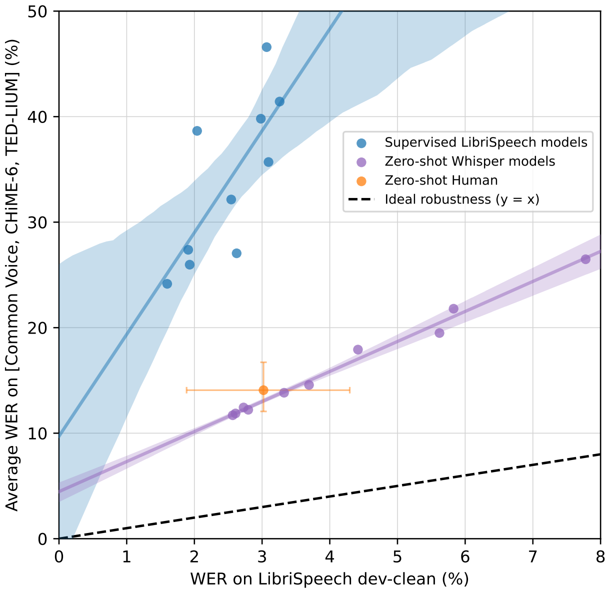

Figure 2. **Zero-shot Whisper models close the gap to human robustness.** Despite matching or outperforming a human on LibriSpeech dev-clean, supervised LibriSpeech models make roughly twice as many errors as a human on other datasets demonstrating their brittleness and lack of robustness. The estimated robustness frontier of zero-shot Whisper models, however, includes the 95% confidence interval for this particular human.

Our main findings are summarized in Figure 2 and Table 1. Although the best zero-shot Whisper model has a relatively unremarkable LibriSpeech clean-test WER of 2.5, which is roughly the performance of modern supervised baseline or the mid-2019 state of the art, zero-shot Whisper models have very different robustness properties than supervised LibriSpeech models and out-perform all benchmarked LibriSpeech models by large amounts on other datasets. Even the smallest zero-shot Whisper model, which has only 39 million parameters and a 6.7 WER on LibriSpeech test-clean is roughly competitive with the best supervised LibriSpeech model when evaluated on other datasets. When compared to a human in Figure 2, the best zero-shot Whisper models roughly match their accuracy and robustness. For a detailed breakdown of this large improvement in robustness, Table 1 compares the performance of the best zero-shot Whisper model with a supervised LibriSpeech model that has the closest performance to it on LibriSpeech test-clean. Despite their very close performance on the reference distribution, the zero-shot Whisper model achieves an average relative error reduction of 55.2% when evaluated on other speech recognition datasets.

Dataset	wav2vec 2.0 Large (no LM)	Whisper Large V2	RER (%)
LibriSpeech Clean	**2.7**	**2.7**	0.0
Artie	24.5	**6.2**	74.7
Common Voice	29.9	**9.0**	69.9
Fleurs En	14.6	**4.4**	69.9
Tedlium	10.5	**4.0**	61.9
CHiME6	65.8	**25.5**	61.2
VoxPopuli En	17.9	**7.3**	59.2
CORAAL	35.6	**16.2**	54.5
AMI IHM	37.0	**16.9**	54.3
Switchboard	28.3	**13.8**	51.2
CallHome	34.8	**17.6**	49.4
WSJ	7.7	**3.9**	49.4
AMI SDM1	67.6	**36.4**	46.2
LibriSpeech Other	6.2	**5.2**	16.1
Average	29.3	**12.8**	55.2

Table 1. **Detailed comparison of effective robustness across various datasets.** Although both models perform within 0.1% of each other on LibriSpeech, a zero-shot Whisper model performs much better on other datasets than expected for its LibriSpeech performance and makes 55.2% less errors on average. Results reported in word error rate (WER) for both models after applying our text normalizer.

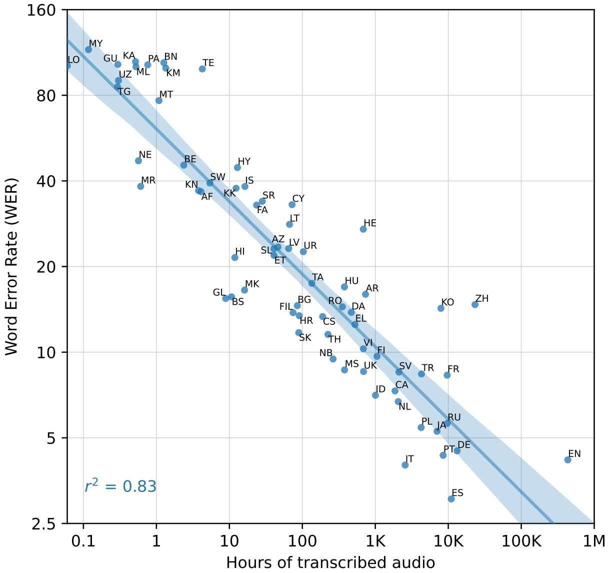

Figure 3. **Correlation of pre-training supervision amount with downstream speech recognition performance.** The amount of pre-training speech recognition data for a given language is very predictive of zero-shot performance on that language in Fleurs.

Model	MLS	VoxPopuli
VP-10K + FT	-	15.3
XLS-R (1B)	10.9	10.6
mSLAM-CTC (2B)	9.7	9.1
Maestro	-	**8.1**
Zero-Shot Whisper	**7.3**	13.6

Table 2. **Multilingual speech recognition performance.** Zero-shot Whisper improves performance on Multilingual LibriSpeech (MLS) but is still significantly behind both Maestro, XLS-R, and mSLAM on VoxPopuli.

This finding suggests emphasizing zero-shot and out-of-distribution evaluations of models, particularly when attempting to compare to human performance, to avoid overstating the capabilities of machine learning systems due to misleading comparisons.

3.4. Multi-lingual Speech Recognition

In order to compare to prior work on multilingual speech recognition, we report results on two low-data benchmarks: Multilingual LibriSpeech (MLS) (Pratap et al., 2020b) and VoxPopuli (Wang et al., 2021) in Table 2.

Whisper performs well on Multilingual LibriSpeech, outperforming XLS-R (Babu et al., 2021), mSLAM (Bapna et al., 2022), and Maestro (Chen et al., 2022b) in a zero-shot setting. We caution that we do use a simple text standardizer for this result which prevents direct comparison or claims of SOTA performance. On VoxPopuli, however, Whisper significantly underperforms prior work and only beats the VP-10K+FT baseline from the original paper. We suspect the underperformance of Whisper models on VoxPopuli could be due to other models including this distribution as a major source for their unsupervised pre-training data and the dataset having significantly more supervised data, which benefits fine-tuning. While MLS has 10 hours of training data per language, the average amount of training data per language is roughly 10× higher for VoxPopuli.

To study the performance of Whisper more broadly we also report performance on the Fleurs dataset (Conneau et al., 2022). In particular, we were interested in studying the relationship between the amount of training data we have for a given language and the resulting downstream zero-shot performance for that language. We visualize this relation in Figure 3. We find a strong squared correlation coefficient of 0.83 between the log of the word error rate and the log of the amount of training data per language. Checking the regression coefficient for a linear fit to these log-log values results in an estimate that WER halves for every 16× increase in training data. We also observed that many of the largest outliers in terms of worse than expected performance according to this trend are languages that have unique scripts and are more distantly related to the Indo-European languages making up the majority of the training dataset such as Hebrew (HE), Telugu (TE), Chinese

X → English	High	Mid	Low	All
XMEF-X	34.2	20.2	5.9	14.7
XLS-R (2B)	36.1	27.7	15.1	22.1
mSLAM-CTC (2B)	37.8	29.6	18.5	24.8
Maestro	**38.2**	31.3	18.4	25.2
Zero-Shot Whisper	36.2	**32.6**	**25.2**	**29.1**

Table 3. **X→en Speech translation performance.** Zero-shot Whisper outperforms existing models on CoVoST2 in the overall, medium, and low resource settings but still moderately underperforms on high-resource languages compared to prior directly supervised work.

(ZH), and Korean (KO). These differences could be due to a lack of transfer due to linguistic distance, our byte level BPE tokenizer being a poor match for these languages, or variations in data quality.

3.5. Translation

We study the translation capabilities of Whisper models by measuring their performance on the X→en subset of CoVoST2 (Wang et al., 2020b). We compare with Maestro, mSLAM, and XLS-R, the highest-performing prior work. We achieve a new state of the art of 29.1 BLEU zero-shot without using any of the CoVoST2 training data. We attribute this to the 68,000 hours of X→en translation data for these languages in our pre-training dataset which, although noisy, is vastly larger than the 861 hours of training data for X→en translation in CoVoST2. Since Whisper evaluation is zero-shot, it does particularly well on the lowest resource grouping of CoVoST2, improving over mSLAM by 6.7 BLEU. Conversely, the best Whisper model does not actually improve over Maestro and mSLAM on average for the highest resource languages.

3.6. Language Identification

To evaluate language identification, we use the Fleurs dataset (Conneau et al., 2022). The zero-shot performance of Whisper AT 64.5% is not competitive with prior supervised work here and underperforms the supervised SOTA of 77.7% BY mSLAM-CTC 2B by 13.6%. Whisper is heavily disadvantaged for language identification on Fleurs, since the Whisper dataset contains no training data for 20 of the 102 languages in Fleurs, upper-bounding accuracy at 80.4%. On the 82 overlapping languages the best Whisper model achieves 80.3% accuracy.

Dataset size	English WER (↓)	Multilingual WER (↓)	X→En BLEU (↑)
3405	30.5	92.4	0.2
6811	19.6	72.7	1.7
13621	14.4	56.6	7.9
27243	12.3	45.0	13.9
54486	10.9	36.4	19.2
681070	**9.9**	**29.2**	**24.8**

Table 4. **Performance improves with increasing dataset size.** English speech recognition performance refers to an average over 12 datasets while the Multilingual speech recognition reports performance on the overlapping subset of languages in Fleurs and X→en translation reports average BLEU on CoVoST2. Dataset size reported in hours.

4. Analysis and Ablations

4.1. Dataset Scaling

At 680,000 hours of labeled audio, the Whisper dataset is one of the largest created in supervised speech recognition. To measure how important dataset size is to Whisper's performance, we trained a series of medium-sized models on subsampled versions of the dataset and compared their performance with the same medium-sized model trained on the whole dataset. Early stopping based on the validation loss was used to select model checkpoints for each dataset size. Evaluation was performed on an exponential moving average estimate of the parameters (Polyak & Juditsky, 1992) using a smoothing rate of 0.9999 to help reduce the effect of the learning rate not fully decaying to zero for the models trained on the subsampled datasets due to early stopping. Performance on English and multilingual speech recognition and X→en translation is reported in Table 4.

All increases in the dataset size result in improved performance on all tasks, although we see significant variability in improvement rates across tasks and sizes. Performance improves rapidly on English speech recognition from 3,000 to 13,000 hours and then slows down noticeably between 13,000 and 54,000 hours. Using the full dataset, which corresponds to another 12.5× increase in size results in only a further 1 point drop in WER. This mirrors the diminishing returns observed with model size scaling for English speech recognition and could similarly be explained by saturation effects when approaching human-level performance.

Improvements in WER scale smoothly for multilingual speech recognition till 54,000 hours and then diminish, improving only a further 7 points when increasing to the full dataset size. For X→en translation, performance is almost zero when training on 7,000 hours of audio or less, and then follows a roughly log-linear improvement till 54,000 hours before also showing diminishing returns when further scaling to the full dataset size.

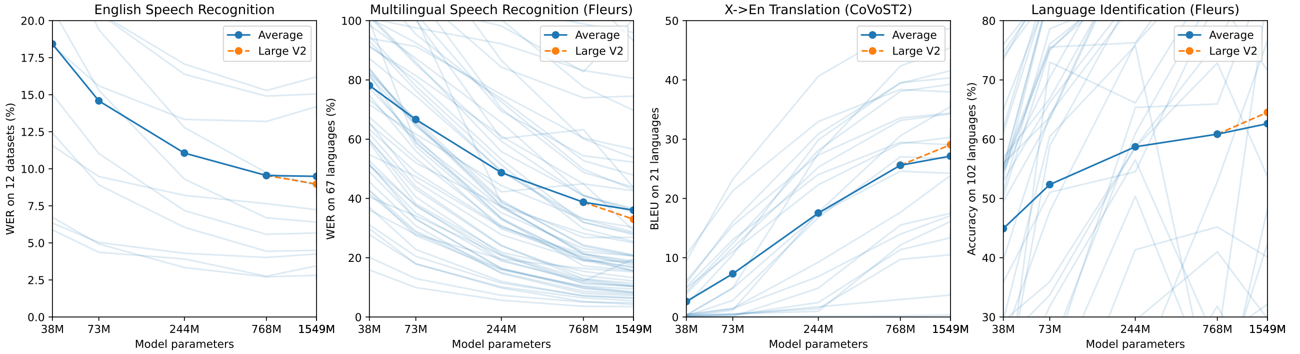

Figure 4. **Zero-shot Whisper performance scales reliably across tasks and languages with increasing model size.** Lightly shaded lines represent individual datasets or languages, showing that performance is more varied than the smooth trends in aggregate performance. Large V2 distinguished with a dashed orange line since it includes several changes that are not present for the smaller models in this analysis.

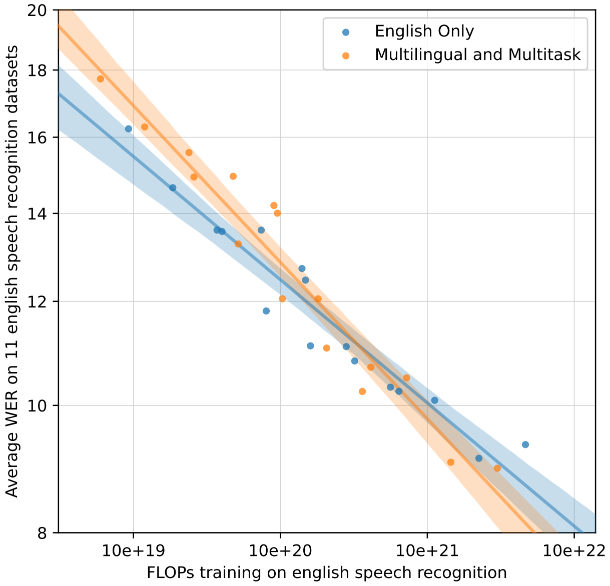

Figure 5. **Multitask and multilingual transfer improves with scale** and eventually outperform models trained on English data only. 95% bootstrap estimate confidence intervals are shown.

4.2. Model Scaling

We also study the zero-shot generalization of Whisper models as a function of the model size. Our analysis is summarized in Figure 4. With the exception of English speech recognition, performance continues to increase with model size across multilingual speech recognition, speech translation, and language identification. The diminishing returns for English speech recognition could be due to saturation effects from approaching human-level performance as analysis in Appendix Section A.3 suggests.

4.3. Multitask and Multilingual Transfer

A concern with jointly training a single model is interference between tasks and languages which could result in performance worse than single task or language models. To investigate whether this is occurring, we compared the performance of models trained on just English speech recognition with our multitask and multilingual training setup and measured their average performance across our suite of zero-shot English speech recognition benchmarks. We adjust for the amount of training on the task of English speech recognition as only 65% of compute is spent on this task in a joint setup; analysis would otherwise be confounded by under-training on the task when compared to a same-sized English-only model.

Our results visualized in Figure 5 show that for small models trained with moderate amounts of compute, there is indeed negative transfer between tasks and languages: joint models underperform English-only models trained for the same amount of compute. However, multitask and multilingual models scale better and for our largest models outperform English-only training demonstrating positive transfer from other tasks.

5. Related Work

Scaling Speech Recognition Early work applying deep learning to speech recognition found improved performance with model depth and size and leveraged GPU acceleration to make training larger models tractable (Mohamed et al., 2009). Further research demonstrated that the benefit of deep learning approaches to speech recognition increased with dataset size, improving from being only competitive with prior GMM-HMM systems when using just 3 hours of TIMIT training data for phone recognition to achieving a 30% word error rate reduction when trained on the 2,000 hour Switchboard dataset (Seide et al., 2011). Liao et al. (2013) is an early example of leveraging weakly supervised learning to increase the size of a deep learning based speech recognition dataset by over 1,000 hours. These trends continued with Deep Speech 2 (Amodei et al., 2015) being a notable system developing high-throughput distributed training across 16 GPUs and scaling to 12,000 hours of training data while demonstrating continuing improvements at that scale. By leveraging semi-supervised pre-training, Narayanan et al. (2018) were able to grow dataset size much further and study training on 162,000 hours of labeled audio. More recent work has explored billion-parameter models (Zhang et al., 2020) and using up to 1,000,000 hours of training data (Zhang et al., 2021).

Multitask Learning Multitask learning (Caruana, 1997) has been studied for a long time. In speech recognition, multi-lingual models have been explored for well over a decade (Schultz & Kirchhoff, 2006). An inspirational and foundational work in NLP exploring multi-task learning with a single model is Collobert et al. (2011). Multitask learning in the sequence-to-sequence framework (Sutskever et al., 2014) using multiple encoders and decoders was investigated in Luong et al. (2015). The use of language codes with a shared encoder/decoder architecture was first demonstrated for machine translation by Johnson et al. (2017), removing the need for separate encoders and decoders. This approach was simplified further into the "text-to-text" framework of McCann et al. (2018) and popularized by its success with large transformer language models in the work of Radford et al. (2019) and Raffel et al. (2020). Toshniwal et al. (2018) demonstrated jointly training a modern deep learning speech recognition system on several languages with a single model, and Pratap et al. (2020a) scaled this line of work significantly to 50 languages with a billion-parameter model. MUTE (Wang et al., 2020c) and mSLAM (Bapna et al., 2022) studied joint training over both text and speech language tasks, demonstrating transfer between them.

Robustness Torralba & Efros (2011) highlighted the lack of generalization of machine learning models between datasets over a decade ago. Many other works have shown and continually reiterated how despite high performance on IID test sets, machine learning models can still make many mistakes when evaluated in even slightly different settings (Lake et al., 2017; Jia & Liang, 2017; Alcorn et al., 2019; Barbu et al., 2019; Recht et al., 2019). More recently, Taori et al. (2020) studied the robustness of image classification models, and Miller et al. (2020) investigated this for question-answering models. A key finding has been that multi-domain training increases robustness and generalization as discussed in the Introduction. This finding has been replicated across many fields in addition to speech recognition including NLP (Hendrycks et al., 2020) and computer vision (Radford et al., 2021).

6. Conclusion

Whisper suggests that scaling weakly supervised pre-training has been underappreciated so far in speech recognition research. We achieve our results without the need for the self-supervision and self-training techniques that have been a mainstay of recent large-scale speech recognition work and demonstrate how simply training on a large and diverse supervised dataset and focusing on zero-shot transfer can significantly improve the robustness of a speech recognition system.

Acknowledgments

We'd like to thank the millions of people who were involved in creating the data used by Whisper. We'd also like to thank Nick Ryder, Will Zhuk, and Andrew Carr for the conversation on the waterfall hike that inspired this project. We are also grateful to the Acceleration and Supercomputing teams at OpenAI for their critical work on software and hardware infrastructure this project used. We'd also like to thank Pamela Mishkin for advising the project from a policy perspective. Finally, we are grateful to the developers of the many software packages used throughout this project including, but not limited, to Numpy (Harris et al., 2020), SciPy (Virtanen et al., 2020), ftfy (Speer, 2019), PyTorch (Paszke et al., 2019), pandas (pandas development team, 2020), and scikit-learn (Pedregosa et al., 2011).

References

Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.-S., and Nguyen, A. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4845–4854, 2019.

Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., et al. Deep speech 2: end-to-end speech recognition in english and mandarin. arxiv. *arXiv preprint arXiv:1512.02595*, 2015.

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

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., et al. XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*, 2021.

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

Baevski, A., Hsu, W.-N., Conneau, A., and Auli, M. Unsupervised speech recognition. *Advances in Neural Information Processing Systems*, 34:27826–27839, 2021.

Bapna, A., Cherry, C., Zhang, Y., Jia, Y., Johnson, M., Cheng, Y., Khanuja, S., Riesa, J., and Conneau, A. mslam: Massively multilingual joint pre-training for speech and text. *arXiv preprint arXiv:2202.01374*, 2022.

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019.

Caruana, R. Multitask learning. *Machine learning*, 28(1): 41–75, 1997.

Chan, W., Park, D., Lee, C., Zhang, Y., Le, Q., and Norouzi, M. SpeechStew: Simply mix all available speech recognition data to train one large neural network. *arXiv preprint arXiv:2104.02133*, 2021.

Chen, G., Chai, S., Wang, G., Du, J., Zhang, W.-Q., Weng, C., Su, D., Povey, D., Trmal, J., Zhang, J., et al. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*, 2021.

Chen, S., Wu, Y., Wang, C., Chen, Z., Chen, Z., Liu, S., Wu, J., Qian, Y., Wei, F., Li, J., et al. Unispeech-sat: Universal speech representation learning with speaker aware pre-training. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6152–6156. IEEE, 2022a.

Chen, Z., Zhang, Y., Rosenberg, A., Ramabhadran, B., Moreno, P., Bapna, A., and Zen, H. Maestro: Matched speech text representations through modality matching. *arXiv preprint arXiv:2204.03409*, 2022b.

Child, R., Gray, S., Radford, A., and Sutskever, I. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537, 2011.

Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C., and Bapna, A. Fleurs: Few-shot learning evaluation of universal representations of speech. *arXiv preprint arXiv:2205.12446*, 2022.

Del Rio, M., Delworth, N., Westerman, R., Huang, M., Bhandari, N., Palakapilly, J., McNamara, Q., Dong, J., Zelasko, P., and Jetté, M. Earnings-21: a practical benchmark for asr in the wild. *arXiv preprint arXiv:2104.11348*, 2021.

Galvez, D., Diamos, G., Torres, J. M. C., Achorn, K., Gopi, A., Kanter, D., Lam, M., Mazumder, M., and Reddi, V. J. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. *arXiv preprint arXiv:2111.09344*, 2021.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Gunter, K., Vaughn, C., and Kendall, T. Contextualizing/s/retraction: Sibilant variation and change in washington dc african american language. *Language Variation and Change*, 33(3):331–357, 2021.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. Array programming with NumPy. *Nature*, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2.

Hendrycks, D. and Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., and Song, D. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*, 2020.

Hernandez, F., Nguyen, V., Ghannay, S., Tomashenko, N. A., and Estève, Y. Ted-lium 3: twice as much data and corpus repartition for experiments on speaker adaptation. In *SPECOM*, 2018.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021a.

Hsu, W.-N., Sriram, A., Baevski, A., Likhomanenko, T., Xu, Q., Pratap, V., Kahn, J., Lee, A., Collobert, R., Synnaeve, G., et al. Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. *arXiv preprint arXiv:2104.01027*, 2021b.

Huang, G., Sun, Y., Liu, Z., Sedra, D., and Weinberger, K. Q. Deep networks with stochastic depth. In *European conference on computer vision*, pp. 646–661. Springer, 2016.

Jia, R. and Liang, P. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*, 2017.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

Kendall, T. and Farrington, C. The corpus of regional african american language. Version 2021.07. Eugene, OR: The Online Resources for African American Language Project. `http://oraal.uoregon.edu/coraal`, 2021. Accessed: 2022-09-01.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., and Goel, S. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. Big transfer (bit): General visual representation learning. In *European conference on computer vision*, pp. 491–507. Springer, 2020.

Kuchaiev, O., Li, J., Nguyen, H., Hrinchuk, O., Leary, R., Ginsburg, B., Kriman, S., Beliaev, S., Lavrukhin, V., Cook, J., et al. Nemo: a toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*, 2019.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

Liao, H., McDermott, E., and Senior, A. Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 368–373. IEEE, 2013.

Likhomanenko, T., Xu, Q., Pratap, V., Tomasello, P., Kahn, J., Avidov, G., Collobert, R., and Synnaeve, G. Rethinking evaluation in asr: Are our models robust enough? *arXiv preprint arXiv:2010.11745*, 2020.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O., and Kaiser, L. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*, 2015.

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., and Van Der Maaten, L. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 181–196, 2018.

Mauch, M. and Ewert, S. The audio degradation toolbox and its application to robustness evaluation. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, Curitiba, Brazil, 2013. accepted.

McCann, B., Keskar, N. S., Xiong, C., and Socher, R. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*, 2018.

Meyer, J., Rauchenstein, L., Eisenberg, J. D., and Howell, N. Artie bias corpus: An open dataset for detecting demographic bias in speech applications. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 6462–6468, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL `https://aclanthology.org/2020.lrec-1.796`.

Miller, J., Krauth, K., Recht, B., and Schmidt, L. The effect of natural distribution shift on question answering models. In *ICML*, 2020.

Mohamed, A.-r., Dahl, G., Hinton, G., et al. Deep belief networks for phone recognition. In *Nips workshop on deep learning for speech recognition and related applications*, volume 1, pp. 39, 2009.

Narayanan, A., Misra, A., Sim, K. C., Pundak, G., Tripathi, A., Elfeky, M., Haghani, P., Strohman, T., and Bacchiani, M. Toward domain-invariant speech recognition via large scale training. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 441–447. IEEE, 2018.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

pandas development team, T. pandas-dev/pandas: Pandas, February 2020. URL `https://doi.org/10.5281/zenodo.3509134`.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. SpecAugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

Pascanu, R., Mikolov, T., and Bengio, Y. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pp. 1310–1318. PMLR, 2013.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Polyak, B. T. and Juditsky, A. B. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Pratap, V., Sriram, A., Tomasello, P., Hannun, A. Y., Liptchinsky, V., Synnaeve, G., and Collobert, R. Massively multilingual asr: 50 languages, 1 model, 1 billion parameters. *ArXiv*, abs/2007.03001, 2020a.

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., and Collobert, R. Mls: A large-scale multilingual dataset for speech research. *arXiv preprint arXiv:2012.03411*, 2020b.

Press, O. and Wolf, L. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 157–163, Valencia, Spain, April 2017. Association for Computational Linguistics. URL `https://aclanthology.org/E17-2025`.

Provilkov, I., Emelianenko, D., and Voita, E. Bpe-dropout: Simple and effective subword regularization. *arXiv preprint arXiv:1910.13267*, 2019.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R. D., and Bengio, Y. SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do ImageNet classifiers generalize to ImageNet? In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/recht19a.html`.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252, 2015.

Schultz, T. and Kirchhoff, K. *Multilingual speech processing*. Elsevier, 2006.

Seide, F., Li, G., Chen, X., and Yu, D. Feature engineering in context-dependent deep neural networks for conversational speech transcription. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pp. 24–29. IEEE, 2011.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

Speer, R. ftfy. Zenodo, 2019. URL `https://doi.org/10.5281/zenodo.2591652`. Version 5.5.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18583–18599. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/d8330f857a17c53d217014ee776bfd50-Paper.pdf`.

Torralba, A. and Efros, A. A. Unbiased look at dataset bias. *CVPR 2011*, pp. 1521–1528, 2011.

Toshniwal, S., Sainath, T. N., Weiss, R. J., Li, B., Moreno, P. J., Weinstein, E., and Rao, K. Multilingual speech recognition with a single end-to-end model. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4904–4908, 2018.

Valk, J. and Alumäe, T. Voxlingua107: a dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 652–658. IEEE, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Wang, C., Tang, Y., Ma, X., Wu, A., Okhonko, D., and Pino, J. fairseq s2t: Fast speech-to-text modeling with fairseq. *arXiv preprint arXiv:2010.05171*, 2020a.

Wang, C., Wu, A., and Pino, J. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310*, 2020b.

Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., and Dupoux, E. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*, 2021.

Wang, P., Sainath, T. N., and Weiss, R. J. Multitask training with text data for end-to-end speech recognition. *arXiv preprint arXiv:2010.14318*, 2020c.

Watanabe, S., Mandel, M., Barker, J., Vincent, E., Arora, A., Chang, X., Khudanpur, S., Manohar, V., Povey, D., Raj, D., et al. Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. *arXiv preprint arXiv:2004.09249*, 2020.

Xu, Q., Baevski, A., Likhomanenko, T., Tomasello, P., Conneau, A., Collobert, R., Synnaeve, G., and Auli, M. Self-training and pre-training are complementary for speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3030–3034. IEEE, 2021.

Zhang, Y., Qin, J., Park, D. S., Han, W., Chiu, C.-C., Pang, R., Le, Q. V., and Wu, Y. Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2010.10504*, 2020.

Zhang, Y., Park, D. S., Han, W., Qin, J., Gulati, A., Shor, J., Jansen, A., Xu, Y., Huang, Y., Wang, S., et al. BigSSL: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2109.13226*, 2021.

A. Long-form Transcription

A.1. Strategies for Reliable Long-form Transcription

Transcribing long-form audio using Whisper relies on accurate prediction of the timestamp tokens to determine the amount to shift the model's 30-second audio context window by, and inaccurate transcription in one window may negatively impact transcription in the subsequent windows. We have developed a set of heuristics that help avoid failure cases of long-form transcription. First, we use beam search with 5 beams using the log probability as the score function, to reduce repetition looping which happens more frequently in greedy decoding. We start with temperature 0, i.e. always selecting the tokens with the highest probability, and increase the temperature by 0.2 up to 1.0 when either the average log probability over the generated tokens is lower than -1 or the generated text has a gzip compression rate higher than 2.4. Providing the transcribed text from the preceding window as previous-text conditioning when the applied temperature is below 0.5 further improves the performance. We found that the probability of the `<|nospeech|>` token alone is not sufficient to distinguish a segment with no speech, but combining the no-speech probability threshold of 0.6 and the average log-probability threshold of -1 makes the voice activity detection of Whisper more reliable. Finally, to avoid a failure mode where the model ignores the first few words in the input, we constrained the initial timestamp token to be between 0.0 and 1.0 second. Table 5 shows that adding each of the interventions above incrementally reduces the WER overall, but not evenly across the dataset. These heuristics serve as a workaround for the noisy predictions of the model, and more research would be needed to further improve the reliability of long-form decoding.

	TED-LIUM3	Meanwhile	Kincaid46	Rev16	Earnings-21	Earnings-22	CORAAL	Average
Greedy decoding only	3.95	5.16	9.69	11.7	10.7	14.0	22.0	11.0
+ Beam search	4.16	5.71	9.42	11.5	10.2	13.4	20.0	10.6
+ Temperature fallback	4.16	5.71	9.42	11.5	10.2	13.4	20.0	10.6
+ Voice activity detection	3.56	**4.61**	9.45	11.4	10.1	13.2	19.4	10.2
+ Previous text conditioning	**3.42**	6.16	8.72	**11.0**	**9.63**	13.3	**18.1**	10.0
+ Initial timestamp constraint	3.51	5.26	**8.41**	11.5	9.73	**12.6**	19.1	10.0

Table 5. Long-form transcription performance improves incrementally as additional decoding heuristics are employed. Details on each intervention are described in Section A.1.

A.2. Comparison with Other ASR Models

Whisper models are trained on 30-second audio chunks and cannot consume longer audio inputs at once. This is not a problem with most academic datasets comprised of short utterances but presents challenges in real-world applications which often require transcribing minutes- or hours-long audio. We developed a strategy to perform buffered transcription of long audio by consecutively transcribing 30-second segments of audio and shifting the window according to the timestamps predicted by the model. We observed that it is crucial to have beam search and temperature scheduling based on the repetitiveness and the log probability of the model predictions in order to reliably transcribe long audio. The full procedure is described in Section A.1.

We evaluate the long-form transcription performance on seven datasets consisting of speech recordings of various lengths and recording conditions, to cover as diverse a data distribution as possible. These include a long-form adaptation of TED-LIUM3 (Hernandez et al., 2018) concatenated so that each example is a full-length TED talk, a collection of jargon-laden segments taken from The Late Show with Stephen Colbert (Meanwhile), sets of videos/podcasts that has been used as ASR benchmarks in online blogs (Rev16 and Kincaid46), recordings of earnings calls (Del Rio et al., 2021), and the full-length interviews from the Corpus of Regional African American Language (CORAAL) (Gunter et al., 2021). Full details about the long-form datasets can be found in Appendix B.

We compare the performance with open-source models as well as 4 commercial ASR services. The results are summarized in Figure 6, showing the distribution of word error rates from Whisper and the 4 commercial ASR services, as well as the NVIDIA STT Conformer-CTC Large model from the NeMo toolkit (Kuchaiev et al., 2019) which performed the best among the open-source models. All commercial ASR services are queried using their default English transcription settings as of September 1st, 2022, and for the NVIDIA STT model we used their buffered inference implementation in the `FrameBatchASR` class to enable long-form transcription. The results show that Whisper performs better than the compared models on most datasets, especially on the Meanwhile dataset which is heavy with uncommon words. Additionally, we note the possibility that some of the commercial ASR systems have been trained on some of these publicly available datasets, and therefore these results may not be accurately reflecting the relative robustness of the systems.

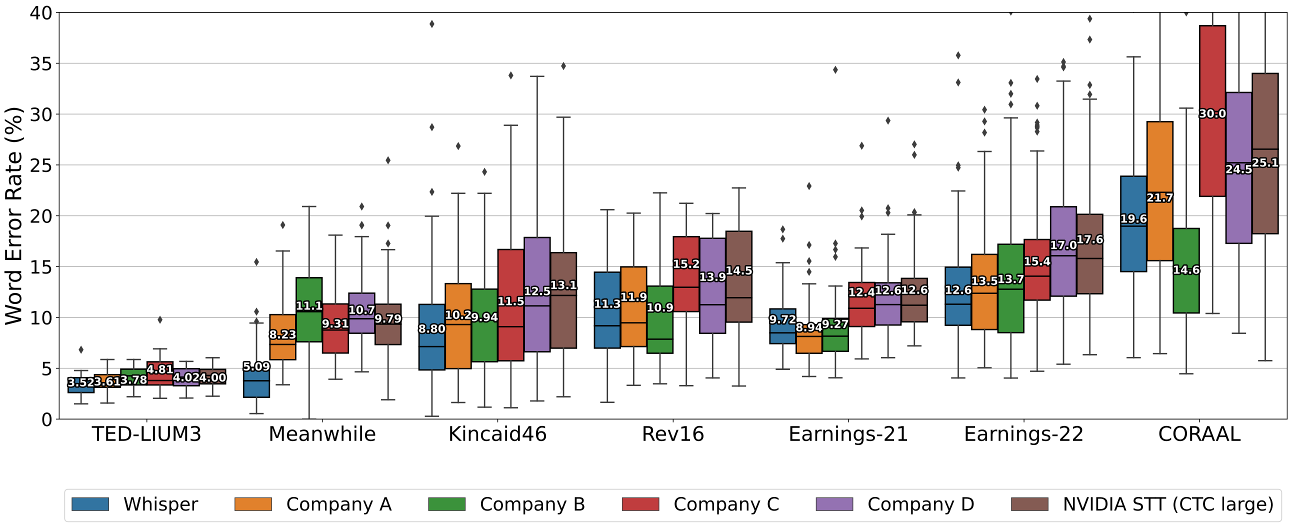

Figure 6. **Whisper is competitive with state-of-the-art commercial and open-source ASR systems in long-form transcription.** The distribution of word error rates from six ASR systems on seven long-form datasets are compared, where the input lengths range from a few minutes to a few hours. The boxes show the quartiles of per-example WERs, and the per-dataset aggregate WERs are annotated on each box. Our model outperforms the best open source model (NVIDIA STT) on all datasets, and in most cases, commercial ASR systems as well.

A.3. Comparison with Human Performance

Because of ambiguous or indistinct speech as well as labeling errors, there are different levels of irreducible error in each dataset, and with WER metrics from ASR systems alone it is difficult to make sense of how much room for improvement exists in each dataset. To quantify how close Whisper's performance is to the human performance, we selected 25 recordings from the Kincaid46 dataset and used 5 services to obtain transcripts produced by professional transcribers, among which one provides computer-assisted transcription and the other four are entirely human-transcribed. The audio selection covers various recording conditions such as scripted and unscripted broadcast, telephone and VoIP calls, and meetings. Figure 7 shows the distribution of per-example WERs and aggregate WER across the 25 recordings, where the computer-assisted service has the lowest aggregate WER that is 1.15% point better than Whisper's, and the pure-human performance is only a fraction of a percentage point better than Whisper's. These results indicate that Whisper's English ASR performance is not perfect but very close to human-level accuracy.

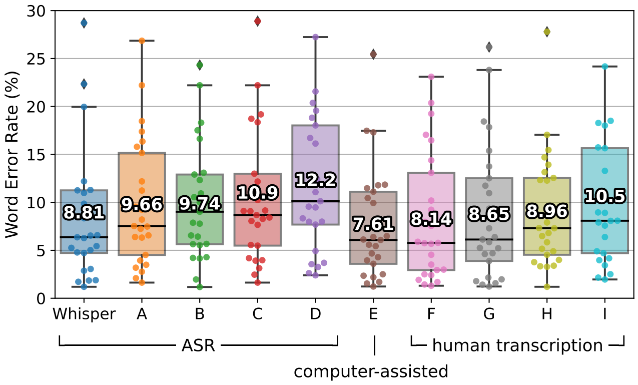

Figure 7. **Whisper's performance is close to that of professional human transcribers.** This plot shows the WER distributions of 25 recordings from the Kincaid46 dataset transcribed by Whisper, the same 4 commercial ASR systems from Figure 6 (A-D), one computer-assisted human transcription service (E) and 4 human transcription services (F-I). The box plot is superimposed with dots indicating the WERs on individual recordings, and the aggregate WER over the 25 recordings are annotated on each box.

B. Evaluation Datasets.

B.1. Short-form English-only datasets

- **LibriSpeech** (Panayotov et al., 2015): We used the test-clean and test-other splits from the LibriSpeech ASR corpus.
- **TED-LIUM 3** (Hernandez et al., 2018): We used the test split of TED-LIUM Release 3, using the segmented manual transcripts included in the release.
- **Common Voice 5.1** (Ardila et al., 2019): We downloaded the English subset of Common Voice Corpus 5.1 from the official website.
- **Artie bias corpus** (Meyer et al., 2020): We used the Artie bias corpus. This is a subset of the Common Voice dataset.
- **CallHome** and **Switchboard**: We used the two corpora from LDC2002S09 and LDC2002T43.
- **WSJ**: We used LDC93S6B and LDC94S13B and followed the `s5` recipe to preprocess the dataset.
- **CORAAL**: We used the 231 interviews from CORAAL (Kendall & Farrington, 2021) and used the preprocessing script from the FairSpeech project.
- **CHiME-6**: For CHiME-6 (Watanabe et al., 2020), we downloaded the CHiME-5 dataset and followed the stage 0 of the `s5_track1` recipe to create the CHiME-6 dataset which fixes synchronization. We then used the binaural recordings (`*_P??.wav`) and the corresponding transcripts.
- **AMI-IHM** and **AMI-SDM1**: We preprocessed the AMI Corpus by following the stage 0 ad 2 of the s5b recipe.

B.2. Long-form English-only datasets

- **TED-LIUM 3** (Hernandez et al., 2018): We used the 11 full-length TED talks from the test split of TED-LIUM Release 3, slicing the source audio files between the beginning of the first labeled segment and the end of the last labeled segment of each talk, and we used the concatenated text as the label.
- **Meanwhile**: This dataset consists of 64 segments from The Late Show with Stephen Colbert. The YouTube video ID and the corresponding start and end timestamps are available as part of the code release. The labels are collected from the closed-caption data for each video and corrected with manual inspection.
- **Rev16**: We use a subset of 16 files from the 30 podcast episodes in Rev.AI's Podcast Transcription Benchmark, after finding that there are multiple cases where a significant portion of the audio and the labels did not match, mostly on the parts introducing the sponsors. We selected 16 episodes that do not have this error, whose "file number"s are:

```
3 4 9 10 11 14 17 18 20 21 23 24 26 27 29 32
```

- **Kincaid46**: This dataset consists of 46 audio files and the corresponding transcripts compiled in the blog article ¡Which automatic transcription service is the most accurate - 2018¿ by Jason Kincaid. We used the 46 audio files and reference transcripts from the Airtable widget in the article. For the human transcription benchmark in the paper, we use a subset of 25 examples from this data, whose "Ref ID"s are:

```
2 4 5 8 9 10 12 13 14 16 19 21 23 25 26 28 29 30 33 35 36 37 42 43 45
```

- **Earnings-21** (Del Rio et al., 2021) and **Earnings-22**: We used the files available in the speech-datasets repository, as of their 202206 version.
- **CORAAL**: We used the 231 full-length interviews and transcripts from (Kendall & Farrington, 2021).

B.3. Multilingual datasets

- **Multilingual LibriSpeech** (Pratap et al., 2020b): We used the test splits from each language in the Multilingual LibriSpeech (MLS) corpus.
- **Fleurs** (Conneau et al., 2022): We collected audio files and transcripts using the implementation available as HuggingFace datasets. To use as a translation dataset, we matched the numerical utterance IDs to find the corresponding transcript in English.
- **VoxPopuli** (Wang et al., 2021): We used the `get_asr_data.py` script from the official repository to collect the ASR data in 16 languages, including English.
- **Common Voice 9** (Ardila et al., 2019): We downloaded the Common Voice Corpus 9 from the official website.
- **CoVOST 2** (Wang et al., 2020b): We collected the X into English data collected using the official repository.

C. Compared Models

For comparison, we use the following models from HuggingFace, downloaded as of September 2022 using version 4.21.0 of the `transformers` library:

- `facebook/wav2vec2-large-960h-lv60-self` (Xu et al., 2021)
- `facebook/wav2vec2-large-robust-ft-libri-960h` (Hsu et al., 2021b)
- `facebook/wav2vec2-base-100h` (Baevski et al., 2020)
- `facebook/wav2vec2-base-960h` (Baevski et al., 2020)
- `facebook/wav2vec2-large-960h` (Baevski et al., 2020)
- `facebook/hubert-large-ls960-ft` (Hsu et al., 2021a)
- `facebook/hubert-xlarge-ls960-ft` (Hsu et al., 2021a)
- `facebook/s2t-medium-librispeech-asr` (Wang et al., 2020a)
- `facebook/s2t-large-librispeech-asr` (Wang et al., 2020a)
- `microsoft/unispeech-sat-base-100h-libri-ft` (Chen et al., 2022a)
- `nvidia/stt_en_conformer_ctc_large` (Kuchaiev et al., 2019)
- `nvidia/stt_en_conformer_transducer_xlarge` (Kuchaiev et al., 2019)
- `speechbrain/asr-crdnn-rnnlm-librispeech` (Ravanelli et al., 2021)
- `speechbrain/asr-transformer-transformerlm-librispeech` (Ravanelli et al., 2021)

We note that all of the models above are entirely or partly trained on LibriSpeech.

D. Text Standardization

D.1. Text Normalization

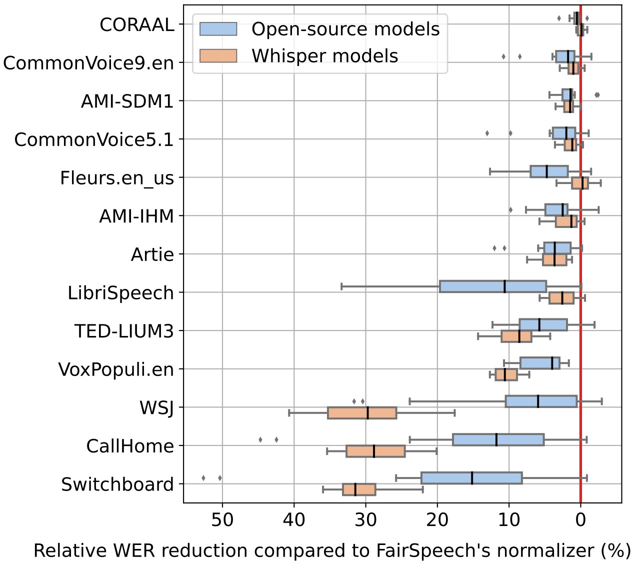

Figure 8. **On most datasets, our text normalizer has similar effect on reducing WERs between Whisper models and other open-source models, compared to FairSpeech's normalizer.** For each dataset, the boxplot shows the distribution of relative WER reduction across different models in our eval suite, showing that using our text normalizer generally results in lower WERs than FairSpeech's. On a few datasets our normalizer reduces WER significantly and more so for Whisper models, such as CallHome and Switchboard which have many contractions in the ground truth and WSJ which contains many numerical expressions.

Since we developed our text normalization jointly with Whisper to discount innocuous word errors, there is a risk that our normalizer is overfitted to fixing Whisper's peculiarities rather than addressing general variation in transcription. To check this, we compared the performance of Whisper using our normalizer versus an independently developed one from the FairSpeech project (Koenecke et al., 2020). In Figure 8, we visualize the differences. On most datasets the two normalizers perform similarly, without significant differences in WER reduction between Whisper and compared open-source models, while on some datasets, namely WSJ, CallHome, and Switchboard, our normalizer reduces the WER of Whisper models' significantly more. The differences in reduction can be traced down to different formats used by the ground truth and how the two normalizers are penalizing them. For example, in CallHome and Switchboard, our standardizer did not penalize differences in common English contractions such as "you're" versus "you are", and in WSJ, our normalizer standardized the written and spoken forms of numerical and monetary expressions, such as "sixty-eight million dollars" versus "$68 million".

Whisper may output any UTF-8 string rather than a restricted set of graphemes, so the rules for text standardization need to be more intricate and comprehensive than those defined on e.g. ASCII characters. We perform the following steps to normalize English texts in different styles into a standardized form, which is a best-effort attempt to penalize only when a word error is caused by actually mistranscribing a word, and not by formatting or punctuation differences.

1. Remove any phrases between matching brackets (`[`, `]`).
2. Remove any phrases between matching parentheses (`(`, `)`).
3. Remove any of the following words: `hmm`, `mm`, `mhm`, `mmm`, `uh`, `um`
4. Remove whitespace characters that comes before an apostrophe `'`
5. Convert standard or informal contracted forms of English into the original form.
6. Remove commas (`,`) between digits

7. Remove periods (`.`) not followed by numbers

8. Remove symbols as well as diacritics from the text, where symbols are the characters with the Unicode category starting with `M`, `S`, or `P`, except period, percent, and currency symbols that may be detected in the next step.

9. Detect any numeric expressions of numbers and currencies and replace with a form using Arabic numbers, e.g. "Ten thousand dollars" → "$10000".

10. Convert British spellings into American spellings.

11. Remove remaining symbols that are not part of any numeric expressions.

12. Replace any successive whitespace characters with a space.

A different, language-specific set of transformations would be needed to equivalently normalize non-English text, but due to our lack of linguistic knowledge to build such normalizers for all languages, we resort to the following basic standardization for non-English text:

1. Remove any phrases between matching brackets (`[`, `]`).

2. Remove any phrases between matching parentheses (`(`, `)`).

3. Replace any markers, symbols, and punctuation characters with a space, i.e. when the Unicode category of each character in the NFKC-normalized string starts with `M`, `S`, or `P`.

4. make the text lowercase.

5. replace any successive whitespace characters with a space.

Additionally, we put a space between every letter for the languages that do not use spaces to separate words, namely Chinese, Japanese, Thai, Lao, and Burmese, effectively measuring the character error rate instead.

We note that the above is an imperfect solution, and it will sometimes produce unintended and unexpected outputs. We do not claim that the text format resulting from the above is more "correct" in any measure. Rather, the procedures above are designed to better distinguish between innocuous differences in wording and genuine mistranscriptions. Python code for the standardization procedures above is available as part of our code and model release to facilitate future iterations and improvements on text standardization.

E. Additional Robustness Analysis

E.1. Robustness to Additive Noise

We tested the noise robustness of Whisper models and 14 LibriSpeech-trained models by measuring the WER when either white noise or pub noise from the Audio Degradation Toolbox (Mauch & Ewert, 2013) was added to the audio. The pub noise represents a more natural noisy environment with ambient noise and indistinct chatter typical in a crowded restaurant or a pub. Among the 14 models, twelve are pre-trained and/or fine-tuned on LibriSpeech, and the other two are NVIDIA STT models trained on a mixture dataset similar to prior work like SpeechStew that includes LibriSpeech. The level of additive noise corresponding to a given signal-to-noise ratio (SNR) is calculated based on the signal power of individual examples. Figure 9 shows how the ASR performance degrades as the additive noise becomes more intensive. There are many models that outperform our zero-shot performance under low noise (40 dB SNR), which is unsurprising given those models are trained primarily on LibriSpeech, but all models quickly degrade as the noise becomes more intensive, performing worse than the Whisper model under additive pub noise of SNR below 10 dB. This showcases Whisper's robustness to noise, especially under more natural distribution shifts like the pub noise.

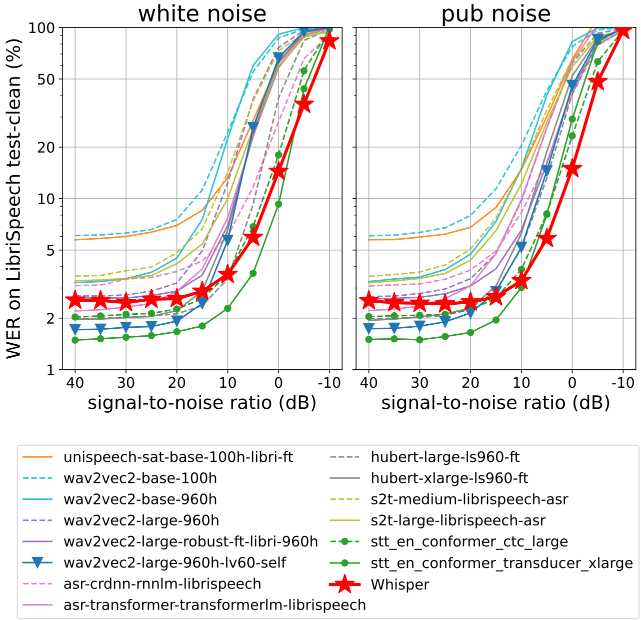

Figure 9. **WER on LibriSpeech test-clean as a function of SNR under additive white noise (left) and pub noise (right).** The accuracy of LibriSpeech-trained models degrade faster than the best Whisper model (★). NVIDIA STT models (●) perform best under low noise but are outperformed by Whisper under high noise (SNR $<$ 10 dB). The second-best model under low noise (▼) is fine-tuned on LibriSpeech only and degrades even more quickly.

F. Raw Performance Tables

F.1. English Transcription

F.1.1. Greedy decoding

Model	LibriSpeech.test-clean	LibriSpeech.test-other	TED-LIUM3	WSJ	CallHome	Switchboard	CommonVoice5.1	Artie	CORAAL	CHiME6	AMI-IHM	AMI-SDM1	VoxPopuli.en	Fleurs.en_us
Whisper tiny.en	5.6	14.6	6.0	5.0	24.1	17.8	26.3	20.0	23.9	41.3	23.7	50.3	11.7	11.6
Whisper tiny	7.6	16.9	7.0	6.7	30.0	22.8	29.6	23.9	31.0	49.6	27.6	58.1	12.7	13.7
Whisper base.en	4.2	10.2	4.9	4.6	20.9	15.2	19.0	13.4	22.6	36.4	20.5	46.7	10.0	7.6
Whisper base	5.0	12.4	5.5	5.1	23.0	16.8	21.6	16.9	26.0	40.2	22.0	49.9	10.0	10.1
Whisper small.en	3.1	7.4	4.0	3.3	18.2	15.7	13.1	9.7	20.2	27.6	17.5	38.0	8.1	6.0
Whisper small	3.4	7.6	4.3	4.0	17.5	14.5	13.5	10.3	18.1	29.3	19.0	39.6	8.3	6.6
Whisper medium.en	3.1	6.3	4.1	3.3	16.2	14.1	10.6	7.6	17.5	25.3	16.4	37.2	7.4	5.0
Whisper medium	2.9	5.9	3.8	2.9	16.4	14.0	10.3	7.2	16.6	26.4	16.6	36.0	7.4	5.4
Whisper large	2.7	5.6	4.0	3.1	15.8	13.1	9.5	6.7	19.4	25.6	16.4	36.9	7.3	4.6
Whisper large-v2	2.7	5.2	4.0	3.9	17.6	13.8	9.0	6.2	16.2	25.5	16.9	36.4	7.3	4.4
wav2vec2-base-100h	6.0	13.4	17.8	13.9	46.9	40.2	47.4	40.8	47.0	79.9	48.1	81.2	28.9	23.1
wav2vec2-base-960h	3.3	8.5	12.8	8.9	40.6	32.9	36.4	30.9	39.9	68.5	40.2	71.9	21.4	17.4
wav2vec2-large-960h-lv60-self	1.8	3.8	7.4	4.4	29.1	22.2	19.9	15.8	29.2	56.3	30.8	57.0	13.0	10.2
wav2vec2-large-960h	2.7	6.2	10.5	7.7	34.8	28.3	29.9	24.5	35.6	65.8	37.0	67.6	17.9	14.6
wav2vec2-large-robust-ft-libri-960h	2.6	5.3	9.2	6.1	23.4	19.8	20.3	16.2	29.4	58.1	31.7	61.6	15.1	11.8
asr-crdnn-rnnlm-librispeech	3.0	9.7	17.7	10.7	59.7	56.1	43.7	33.3	83.8	81.0	57.2	85.8	30.6	32.4
asr-transformer-transformerlm-librispeech	2.1	5.4	11.9	7.4	38.9	33.0	30.6	23.5	44.9	79.5	44.5	75.4	17.8	17.0
hubert-large-ls960-ft	2.0	4.1	8.4	5.4	29.6	22.8	20.8	16.0	32.0	60.0	33.7	59.1	14.4	10.9
hubert-xlarge-ls960-ft	1.9	3.5	8.3	5.4	29.3	22.2	19.8	14.8	31.5	58.5	33.3	58.9	14.2	10.5
s2t-large-librispeech-asr	3.3	8.1	14.9	9.4	54.5	40.3	38.1	30.7	50.2	79.2	53.4	79.5	21.6	18.0
s2t-medium-librispeech-asr	3.6	8.2	15.7	9.7	58.1	42.4	39.3	31.3	52.6	79.8	60.3	85.3	22.9	19.7
stt_en_conformer_ctc_large	2.1	4.2	4.4	2.1	11.3	8.2	7.4	4.0	13.5	30.5	15.9	39.9	6.7	8.2
stt_en_conformer_transducer_xlarge	1.5	2.8	4.3	1.2	12.0	7.4	4.3	1.5	19.9	36.8	20.5	48.6	6.0	6.3
unispeech-sat-base-100h-libri-ft	5.7	13.8	17.7	13.6	46.5	40.0	45.3	38.6	44.7	74.8	47.8	77.7	29.8	22.4

Table 6. English transcription WER (%) with greedy decoding

F.1.2. Beam search with temperature fallback

Model	LibriSpeech.test-clean	LibriSpeech.test-other	TED-LIUM3	WSJ	CallHome	Switchboard	CommonVoice5.1	Artie	CORAAL	CHiME6	AMI-IHM	AMI-SDM1	VoxPopuli.en	Fleurs.en_us
Whisper tiny.en	5.4	12.8	5.4	4.6	21.4	16.0	23.5	18.4	21.4	42.0	22.7	54.2	10.9	10.0
Whisper tiny	6.7	15.0	6.3	5.9	24.8	18.3	26.1	20.8	25.1	48.0	25.6	57.3	11.6	12.4
Whisper base.en	4.1	9.6	4.6	4.0	18.3	14.2	17.5	13.2	18.5	35.2	21.1	49.0	9.3	7.1
Whisper base	4.9	11.0	5.0	4.4	20.5	15.6	19.4	15.3	20.5	40.0	21.5	50.0	9.5	8.9
Whisper small.en	3.2	6.7	4.3	3.0	17.2	13.4	12.6	9.2	17.5	29.5	17.9	42.5	8.1	5.3
Whisper small	3.3	7.2	4.3	3.9	17.1	13.3	12.8	9.3	16.4	30.9	19.2	43.5	8.2	6.1
Whisper medium.en	3.0	5.7	4.3	2.8	14.7	12.4	10.3	7.4	15.3	27.0	17.1	39.4	7.8	4.5
Whisper medium	2.7	5.6	4.0	2.7	15.3	13.2	9.7	6.7	14.9	27.6	17.6	43.0	7.6	4.4
Whisper large	2.8	5.7	4.3	3.5	16.2	14.2	8.9	6.4	15.1	25.2	17.6	37.1	7.2	4.5
Whisper large-v2	2.5	4.9	3.7	2.6	16.4	13.6	8.2	5.7	14.2	24.9	17.4	39.9	7.0	4.2

Table 7. English transcription WER (%) with beam search and temperature fallback

F.2. Multilingual Transcription

F.2.1. MULTILINGUAL LIBRISPEECH

Model	Dutch	English	French	German	Italian	Polish	Portuguese	Spanish
Whisper tiny	39.4	15.7	36.8	24.9	41.7	34.2	31.3	19.2
Whisper base	28.4	11.7	26.6	17.7	31.1	22.8	21.9	12.8
Whisper small	17.2	8.3	16.2	10.5	21.4	11.2	13.0	7.8
Whisper medium	11.7	6.8	8.9	7.4	16.0	6.5	9.0	5.3
Whisper large	10.2	6.3	8.9	6.6	14.3	6.6	9.2	5.4
Whisper large-v2	9.3	6.2	7.3	5.5	13.8	5.0	6.8	4.2

Table 8. WER (%) on MLS

F.2.2. COMMON VOICE 9

Model	Arabic	Bulgarian	Bengali	Catalan	Czech	Welsh	Danish	German	Greek	English	Spanish	Estonian	Persian
Whisper tiny	90.9	79.3	104.1	51.0	79.7	101.8	77.2	34.5	61.9	28.8	30.3	102.1	120.3
Whisper base	84.4	68.1	103.7	39.9	63.1	93.8	57.5	24.5	51.5	21.9	19.6	88.1	99.0
Whisper small	66.4	44.8	118.6	23.8	34.1	65.4	32.1	13.0	31.7	14.5	10.3	67.2	71.9
Whisper medium	60.3	26.7	124.7	16.4	18.8	43.6	19.3	8.5	20.0	11.2	6.9	45.6	49.9
Whisper large	56.0	24.1	106.0	15.3	17.1	40.3	18.3	7.7	18.3	10.1	6.4	41.4	44.8
Whisper large-v2	53.8	19.9	103.4	14.1	13.5	34.2	14.4	6.4	16.0	9.4	5.6	35.1	39.4

Model	Finnish	French	Hindi	Hungarian	Indonesian	Italian	Japanese	Lithuanian	Latvian	Malayalam	Mongolian	Dutch	Polish
Whisper tiny	68.5	49.7	108.3	87.0	49.6	44.5	36.1	103.5	87.8	102.7	123.0	43.6	45.3
Whisper base	52.9	37.3	106.5	71.9	36.1	30.5	24.2	91.3	78.0	122.9	137.0	29.5	32.8
Whisper small	30.5	22.7	43.6	44.4	18.4	16.0	14.0	72.8	54.6	104.8	225.8	14.2	16.9
Whisper medium	18.8	16.0	31.5	26.9	11.6	9.4	10.5	49.4	37.2	137.8	113.4	8.0	10.1
Whisper large	17.0	14.7	25.0	23.5	10.6	8.1	9.4	43.9	34.8	107.1	117.4	7.1	9.0
Whisper large-v2	14.4	13.9	21.9	19.7	8.5	7.1	9.1	35.2	25.5	103.2	128.4	5.8	7.6

Model	Portuguese	Romanian	Russian	Slovak	Slovenian	Serbian	Swedish	Tamil	Thai	Turkish	Urdu	Vietnamese	Chinese
Whisper tiny	35.2	68.2	40.6	104.0	82.0	106.1	58.2	105.7	55.9	53.6	74.7	69.3	52.4
Whisper base	23.7	55.9	28.8	87.2	70.3	103.0	42.4	49.5	32.1	38.6	58.6	51.6	44.9
Whisper small	12.5	33.2	15.0	60.4	45.5	101.3	22.1	28.7	18.1	23.7	39.1	33.3	29.4
Whisper medium	8.1	21.5	9.3	42.0	29.8	85.6	13.7	19.6	10.5	17.7	29.9	24.4	23.2
Whisper large	7.1	19.8	8.2	37.9	25.1	87.4	12.4	17.6	8.8	16.6	28.1	19.9	29.1
Whisper large-v2	6.3	15.8	7.1	31.9	20.6	70.5	10.6	16.1	8.0	14.5	24.2	18.2	26.8

Table 9. WER (%) on CommonVoice9

F.2.3. VOXPOPULI

Model	Czech	German	English	en_accented	Spanish	Estonian	Finnish	French	Croatian	Hungarian	Italian	Lithuanian	Dutch	Polish	Romanian	Slovak	Slovenian
Whisper tiny	73.5	27.4	11.6	18.8	19.7	99.2	54.1	32.9	72.4	74.5	40.5	93.1	41.9	31.4	65.9	78.7	81.9
Whisper base	54.7	20.6	9.5	17.5	14.4	83.0	39.7	24.9	53.6	52.6	30.8	82.1	29.4	22.1	49.3	63.7	70.5
Whisper small	28.8	14.8	8.2	19.2	11.1	59.2	24.9	15.7	33.7	31.3	22.9	60.1	18.8	13.3	28.6	37.3	50.8
Whisper medium	18.4	12.4	7.6	19.1	9.6	38.2	16.6	12.2	23.9	19.3	19.7	39.3	14.9	10.1	18.4	23.0	36.3
Whisper large	15.9	11.9	7.2	20.8	8.8	33.3	15.5	11.0	19.0	16.8	18.4	35.0	14.0	9.0	17.0	19.1	31.3
Whisper large-v2	12.6	11.2	7.0	18.6	8.2	28.7	12.4	11.4	16.1	13.8	19.0	33.2	12.9	7.8	14.4	15.4	27.9

Table 10. WER (%) on VoxPopuli

F.2.4. Fleurs

Model	Afrikaans	Amharic	Arabic	Assamese	Azerbaijani	Belarusian	Bulgarian	Bengali	Bosnian	Catalan	Chinese	Czech	Welsh	Danish
Whisper tiny	91.2	122.9	63.4	102.0	93.1	94.0	81.0	101.6	82.1	42.8	40.5	82.8	101.3	82.0
Whisper base	81.5	196.8	48.8	102.0	76.4	91.3	65.1	100.6	66.7	29.0	34.1	66.0	85.3	57.6
Whisper small	61.1	120.2	30.6	108.0	49.1	75.1	37.3	104.4	39.4	16.2	20.8	37.6	59.3	32.8
Whisper medium	44.9	229.3	20.4	102.3	33.1	60.4	21.4	100.6	23.9	9.6	12.1	21.3	40.8	19.5
Whisper large	42.6	129.3	18.1	105.6	28.7	56.6	18.4	104.9	20.7	8.0	19.6	17.4	36.6	16.8
Whisper large-v2	36.7	140.3	16.0	106.2	23.4	45.4	14.6	104.1	15.7	7.3	14.7	13.3	33.0	13.8

Model	German	Greek	English	Spanish	Estonian	Persian	Finnish	Tagalog	French	Galician	Gujarati	Hausa	Hebrew	Hindi
Whisper tiny	27.8	67.4	12.4	15.9	94.8	101.8	59.5	65.6	41.4	54.8	101.2	100.2	71.6	102.3
Whisper base	17.9	53.5	8.9	9.9	77.9	86.1	43.1	45.8	28.5	47.4	101.4	98.6	61.7	101.1
Whisper small	10.2	30.8	6.1	5.6	51.3	55.8	24.0	27.7	15.0	30.2	106.4	90.1	44.4	38.4
Whisper medium	6.5	19.0	4.4	3.6	29.8	41.0	13.9	19.1	8.7	21.2	104.8	106.6	33.1	26.8
Whisper large	5.5	18.7	4.5	3.5	25.5	36.1	12.2	15.8	7.7	19.0	103.9	87.0	30.2	26.9
Whisper large-v2	4.5	12.5	4.2	3.0	21.9	32.9	9.7	13.8	8.3	15.4	102.7	88.9	27.1	21.5

Model	Croatian	Hungarian	Armenian	Indonesian	Icelandic	Italian	Japanese	Javanese	Georgian	Kazakh	Khmer	Kannada	Korean	Luxembourgish
Whisper tiny	79.0	83.8	118.6	51.7	113.3	29.8	37.0	107.3	123.0	165.2	100.6	100.7	36.1	99.1
Whisper base	59.1	65.0	126.3	33.1	95.5	17.9	22.8	89.5	114.7	109.2	101.6	107.2	27.8	100.7
Whisper small	33.4	38.9	86.6	16.3	72.6	9.8	12.0	88.6	118.3	70.3	104.4	100.4	19.6	100.1
Whisper medium	19.3	24.3	60.1	10.2	49.9	5.2	7.1	67.9	117.3	48.8	98.9	77.7	16.4	90.0
Whisper large	16.7	21.0	53.7	8.5	43.0	4.2	6.4	87.0	100.5	43.8	96.0	69.8	15.2	86.5
Whisper large-v2	13.4	17.0	44.6	7.1	38.2	4.0	5.3	nan	105.0	37.7	99.7	37.0	14.3	88.0

Model	Lingala	Lao	Lithuanian	Latvian	Maori	Macedonian	Malayalam	Mongolian	Marathi	Malay	Maltese	Myanmar	Norwegian	Nepali
Whisper tiny	105.4	115.1	98.5	91.6	94.5	73.3	101.5	113.7	100.3	51.2	100.8	124.8	62.0	101.8
Whisper base	96.7	105.1	87.3	79.8	77.5	59.9	107.4	125.7	100.3	35.1	97.6	122.6	44.0	102.4
Whisper small	91.3	102.2	65.6	53.2	59.5	36.9	100.9	144.2	60.2	18.9	92.2	110.1	24.2	69.5
Whisper medium	83.2	101.4	41.1	32.0	77.8	22.0	101.1	103.7	63.2	12.2	83.2	123.0	12.9	54.4
Whisper large	76.8	101.6	35.2	28.3	45.7	20.6	101.4	106.2	43.7	10.2	80.5	124.5	11.4	52.2
Whisper large-v2	75.6	101.5	28.1	23.1	38.5	16.5	100.7	110.5	38.3	8.7	76.6	115.7	9.5	47.1

Model	Dutch	Occitan	Punjabi	Polish	Pashto	Portuguese	Romanian	Russian	Sindhi	Slovak	Slovenian	Shona	Somali	Serbian
Whisper tiny	49.0	95.9	102.6	45.6	105.6	20.1	74.7	31.1	105.8	77.2	87.2	128.1	105.6	83.7
Whisper base	33.0	82.9	101.5	30.8	99.0	13.0	56.0	20.5	103.9	60.6	74.6	126.0	109.6	64.3
Whisper small	16.4	87.3	103.6	14.7	92.9	7.3	29.8	11.4	131.7	33.3	49.3	140.0	105.3	42.2
Whisper medium	9.9	79.5	102.0	8.0	119.4	5.0	20.0	7.2	147.0	17.3	31.9	143.9	104.0	44.9
Whisper large	8.3	75.9	102.8	7.2	92.7	4.8	15.4	6.4	177.9	15.7	27.8	130.0	103.5	29.2
Whisper large-v2	6.7	75.3	102.4	5.4	93.7	4.3	14.4	5.6	156.5	11.7	23.1	121.0	102.9	33.9

Model	Swedish	Swahili	Tamil	Telugu	Tajik	Thai	Turkish	Ukrainian	Urdu	Uzbek	Vietnamese	Yoruba
Whisper tiny	52.7	100.9	99.9	105.1	101.7	58.8	42.5	51.2	65.2	105.2	60.0	106.4
Whisper base	37.4	92.5	58.7	105.2	109.3	38.2	27.5	37.7	52.0	114.0	40.5	101.8
Whisper small	20.8	73.7	35.2	98.2	84.3	21.9	15.9	19.3	37.3	107.7	21.2	116.4
Whisper medium	11.2	52.8	23.1	82.8	74.0	15.4	10.4	11.6	28.2	109.6	12.7	105.1
Whisper large	10.5	47.9	20.6	100.6	74.5	13.2	9.4	10.3	25.0	93.3	10.7	111.7
Whisper large-v2	8.5	39.3	17.5	99.0	85.8	11.5	8.4	8.6	22.6	90.2	10.3	94.8

Table 11. WER (%) on Fleurs

F.3. Speech Translation

F.3.1. FLEURS

Model	Afrikaans	Amharic	Arabic	Assamese	Azerbaijani	Belarusian	Bulgarian	Bengali	Bosnian	Catalan	Chinese	Czech	Welsh	Danish
Whisper tiny	1.6	0.1	0.1	0.4	0.1	0.8	0.4	0.4	0.4	5.2	0.6	0.6	0.6	0.7
Whisper base	4.4	0.3	1.0	0.4	0.8	3.3	2.7	0.7	4.1	13.1	1.9	2.7	0.7	5.0
Whisper small	18.1	0.2	10.6	1.2	5.8	7.1	14.8	2.7	16.8	25.1	9.3	14.2	1.3	18.1
Whisper medium	29.5	0.9	19.9	3.5	11.7	9.8	23.9	10.6	26.0	31.9	15.1	23.6	8.4	28.6
Whisper large	31.6	1.1	23.8	3.9	13.1	11.0	26.2	12.0	28.0	33.7	16.8	25.6	11.2	31.6
Whisper large-v2	34.1	1.9	25.5	5.4	13.7	11.7	28.5	13.2	29.7	34.2	18.4	27.8	13.0	32.7

Model	German	Greek	English	Spanish	Estonian	Persian	Finnish	Tagalog	French	Galician	Gujarati	Hausa	Hebrew	Hindi
Whisper tiny	5.2	0.1	68.6	7.7	0.1	0.1	0.2	0.8	4.7	4.0	0.7	0.1	0.2	1.0
Whisper base	13.7	0.7	73.3	12.4	0.3	0.2	0.5	2.1	13.1	10.5	1.5	0.0	0.6	3.4
Whisper small	25.9	11.6	77.3	18.2	3.6	5.8	7.3	12.0	23.5	17.5	3.9	0.3	5.4	11.1
Whisper medium	31.4	19.9	79.2	21.4	13.5	15.0	18.5	20.5	28.6	24.7	12.8	0.5	15.9	19.4
Whisper large	34.3	21.7	77.8	22.8	15.9	17.6	20.6	22.7	31.6	26.0	14.8	0.5	19.6	20.7
Whisper large-v2	34.6	23.7	80.2	23.3	18.7	19.6	22.1	24.4	32.2	27.9	16.2	0.4	21.8	22.0

Model	Croatian	Hungarian	Armenian	Indonesian	Icelandic	Italian	Japanese	Javanese	Georgian	Kazakh	Khmer	Kannada	Korean	Luxembourgish
Whisper tiny	0.6	0.1	0.1	0.3	0.4	5.3	0.2	0.2	0.1	0.1	0.1	0.8	0.5	0.8
Whisper base	3.7	0.2	0.1	2.6	0.4	11.3	1.5	0.2	0.2	0.2	0.1	0.9	3.7	1.7
Whisper small	14.6	4.8	0.7	16.4	1.8	17.8	9.6	1.4	0.2	0.8	0.5	2.3	12.2	5.7
Whisper medium	23.0	15.5	10.4	24.1	6.8	21.6	14.9	5.0	1.3	4.3	3.3	8.5	19.2	13.6
Whisper large	25.4	18.3	13.2	27.2	6.6	23.5	17.0	5.1	2.7	6.3	5.2	9.9	20.0	15.4
Whisper large-v2	27.0	21.2	16.0	29.1	9.1	23.6	18.9	6.2	2.4	5.4	6.1	11.6	21.3	16.8

Model	Lingala	Lao	Lithuanian	Latvian	Maori	Macedonian	Malayalam	Mongolian	Marathi	Malay	Maltese	Myanmar	Norwegian	Nepali
Whisper tiny	0.1	0.2	0.1	0.2	0.3	1.0	0.8	0.1	0.2	0.3	0.6	0.1	1.4	0.1
Whisper base	0.1	0.3	0.3	0.4	1.0	5.4	1.4	0.1	0.9	2.1	1.4	0.1	8.4	0.3
Whisper small	0.5	2.0	1.9	1.5	3.9	15.3	5.7	0.1	3.8	14.1	4.9	0.0	22.0	2.9
Whisper medium	0.9	8.1	9.6	10.0	8.5	23.5	13.8	0.5	10.9	23.2	11.2	0.2	29.1	12.7
Whisper large	1.2	9.3	12.0	12.5	9.4	26.4	16.5	1.0	13.1	25.5	12.8	0.5	30.5	12.9
Whisper large-v2	1.0	11.0	14.0	14.3	10.2	27.7	16.7	1.0	12.9	27.3	13.5	0.4	31.4	16.1

Model	Dutch	Occitan	Punjabi	Polish	Pashto	Portuguese	Romanian	Russian	Sindhi	Slovak	Slovenian	Shona	Somali	Serbian
Whisper tiny	2.7	1.7	0.3	0.8	0.3	12.1	1.0	3.1	0.5	0.7	0.3	0.1	0.0	0.6
Whisper base	7.5	4.2	1.1	5.1	0.4	22.4	4.9	12.1	0.7	4.6	1.3	0.3	0.1	5.4
Whisper small	15.9	9.5	4.4	14.0	0.8	31.2	18.3	19.7	2.0	14.4	6.9	0.6	0.1	19.3
Whisper medium	21.6	15.9	12.8	19.0	2.1	35.9	26.6	24.8	5.5	22.7	14.0	1.4	0.4	27.7
Whisper large	22.8	16.8	14.6	21.4	3.7	37.4	29.1	26.7	5.9	25.1	16.9	1.8	0.5	30.5
Whisper large-v2	24.0	20.2	15.7	22.3	3.4	38.1	31.5	27.8	5.7	26.1	17.0	1.8	0.7	32.5

Model	Swedish	Swahili	Tamil	Telugu	Tajik	Thai	Turkish	Ukrainian	Urdu	Uzbek	Vietnamese	Yoruba
Whisper tiny	1.8	0.1	0.2	0.3	0.2	0.2	0.2	1.2	0.4	0.0	0.1	0.2
Whisper base	9.1	0.1	0.4	0.4	0.2	0.7	2.4	6.9	1.5	0.2	0.9	0.5
Whisper small	22.9	0.1	2.1	4.0	4.4	5.8	15.7	18.7	8.8	0.5	8.5	0.5
Whisper medium	32.1	3.1	7.0	10.8	11.4	12.8	22.9	25.8	14.9	3.8	16.6	0.9
Whisper large	33.1	5.3	8.5	10.9	13.0	15.2	25.7	28.0	16.3	5.8	19.5	1.2
Whisper large-v2	35.3	7.2	9.2	12.5	14.5	16.1	26.6	29.4	17.2	6.0	20.4	1.4

Table 12. BLEU scores on Fleurs

F.3.2. CoVoST 2

Model	Arabic	Catalan	Welsh	German	Spanish	Estonian	Persian	French	Indonesian	Italian	Japanese	Latvian	Mongolian	Dutch	Portuguese	Russian	Slovenian	Swedish	Tamil	Turkish	Chinese
Whisper tiny	0.2	4.9	0.4	4.0	10.5	0.2	0.1	6.1	0.3	5.1	0.3	0.1	0.1	4.3	9.5	5.7	0.4	2.0	0.1	0.2	0.4
Whisper base	1.2	11.0	0.5	11.7	21.3	0.3	0.1	15.4	4.9	13.0	4.9	0.5	0.1	12.4	23.2	16.1	1.4	10.5	0.4	2.8	1.4
Whisper small	17.7	22.3	1.0	25.3	33.0	2.4	4.9	27.3	27.6	24.0	17.3	1.4	0.2	28.1	40.6	30.9	9.2	29.9	1.7	16.8	6.8
Whisper medium	30.6	29.2	12.1	33.2	38.4	11.4	15.5	33.6	42.3	29.5	24.6	9.7	0.2	38.1	48.7	39.4	17.7	39.5	2.9	27.0	14.0
Whisper large	35.5	30.3	16.1	34.3	38.0	13.4	17.5	34.4	45.4	29.1	24.2	10.5	0.3	39.3	48.6	41.6	23.9	40.3	3.7	26.7	17.1
Whisper large-v2	39.7	31.8	21.5	36.3	40.1	15.0	19.3	36.4	48.1	30.9	26.1	13.9	0.1	41.2	51.6	43.3	21.6	42.9	4.2	28.3	18.0

Table 13. BLEU scores on CoVoST2

F.4. Long-form Transcription

Model	TED-LIUM3	Meanwhile	Kincaid46	Rev16	Earnings-21	Earnings-22	CORAAL
Whisper tiny.en	5.5	12.8	13.8	15.1	17.0	22.0	30.3
Whisper tiny	6.8	15.5	16.7	17.0	18.7	24.4	33.1
Whisper base.en	4.6	9.4	11.2	13.2	12.5	16.6	25.2
Whisper base	4.8	12.2	12.2	14.5	13.5	18.4	26.9
Whisper small.en	4.6	6.0	9.4	12.0	10.8	14.0	21.9
Whisper small	4.2	6.9	10.1	12.1	11.1	14.3	22.3
Whisper medium.en	3.6	5.2	8.9	11.9	10.2	13.3	20.6
Whisper medium	3.8	5.4	8.6	11.4	10.3	13.2	20.3
Whisper large	3.8	5.3	8.8	11.0	10.3	13.4	20.4
Whisper large-v2	3.5	5.1	8.8	11.3	9.7	12.6	19.6
wav2vec2-base-100h	17.6	27.7	39.3	35.2	45.7	57.1	55.4
wav2vec2-base-960h	12.8	19.7	32.9	29.8	37.3	46.8	49.1
wav2vec2-large-960h-lv60-self	7.2	11.4	21.1	21.3	21.7	28.0	36.7
wav2vec2-large-960h	10.1	16.4	27.4	26.4	30.4	40.1	43.5
wav2vec2-large-robust-ft-libri-960h	8.8	15.2	22.9	23.4	23.0	31.0	36.8
hubert-large-ls960-ft	8.1	12.9	22.4	23.4	23.0	30.6	37.9
hubert-xlarge-ls960-ft	8.1	12.5	22.9	23.2	23.1	31.3	38.1
stt_en_conformer_ctc_large	4.0	9.8	13.1	14.5	12.6	17.6	25.1
stt_en_conformer_transducer_xlarge	5.3	10.6	17.1	19.8	16.2	19.7	38.9

Table 14. Long-form English transcription WER (%)

G. Training Dataset Statistics

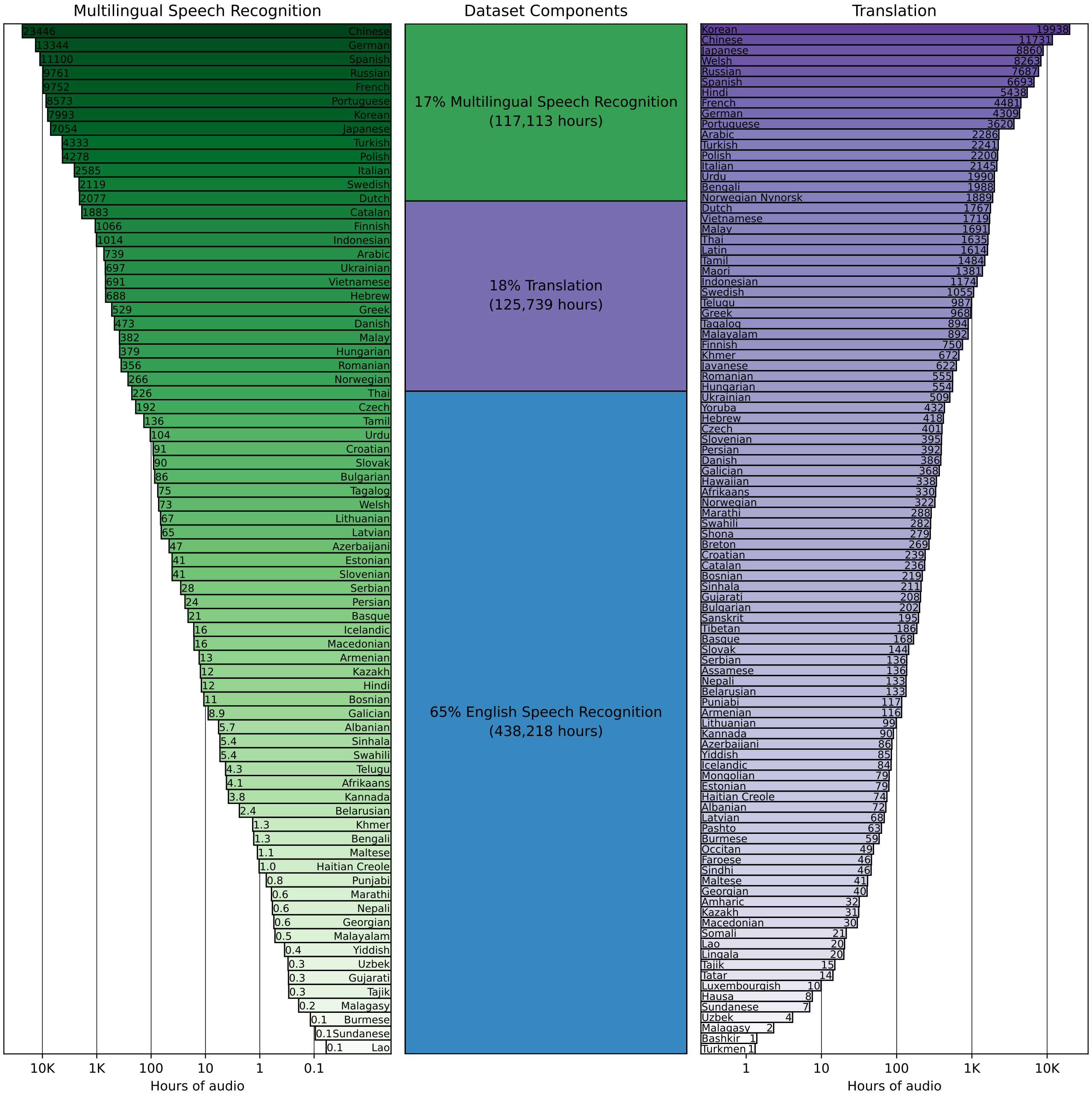

Figure 10. Training dataset statistics

H. Hyperparameters

Model	Layers	Width	Heads	Parameters
Tiny	4	384	6	39M
Base	6	512	8	74M
Small	12	768	12	244M
Medium	24	1024	16	769M
Large	32	1280	20	1550M

Table 15. Architecture details of the Whisper model family.

Hyperparameter	Value
Updates	1048576
Batch Size	256
Warmup Updates	2048
Max grad norm	1.0
Optimizer	AdamW
β_1	0.9
β_2	0.98
ϵ	10^{-6}
Weight Decay	0.1
Weight Init	Gaussian Fan-In
Learning Rate Schedule	Linear Decay
Speechless audio subsample factor	10×
Condition on prior text rate	50%

Table 16. Whisper training hyperparameters.

Hyperparameter	Value
Updates	655360
Batch Size	1024
BPE Dropout	0.1
Stochastic Depth	0.1
SpecAugment Policy	LibriSpeech Basic

Table 17. Hyperparameters changed for Whisper Large V2.

Model	Max Learning Rate
Tiny	1.5×10^{-3}
Base	1×10^{-3}
Small	5×10^{-4}
Medium	2.5×10^{-4}
Large	1.75×10^{-4}
Large V2	2.0×10^{-4}

Table 18. Whisper model learning rates.