Fayoum University Faculty of Engineering, 2025, Vol: 8(1), 131-142
https://fuje.journals.ekb.eg/
ISSN Online: 2537-0634
ISSN Print: 2537-0626

Comparative Study on End-to-End Speech Recognition Using Pre-trained Models

Martha F. Ghobrial[1]*, Amr M. Gody[2],Sayed T. Muhammad[3]

1 Electrical Engineering Department, Faculty of Engineering, Fayoum University, Fayoum, Egypt
2 Professor of Digital Signals, Faculty of Engineering, Fayoum University, Fayoum ,Egypt
3 Assistant Professor, Computers and Systems Engineering Department, Faculty of Engineering, Fayoum University, Fayoum ,Egypt

*Corresponding author: Martha F. Ghobrial, Email: marthafikryghobrial@gmail.com

How to cite this paper: Ghobrial, M.F., Gody, A.M. and Muhammad, S.T. (2025). Comparative Study on End-to-End Speech Recognition Using Pre-trained Models. *Fayoum University Journal of Engineering,* Vol: 8(1), 131-142.
https://dx.doi.org/10.21608/fuje.2024.312102.1089

Abstract

In the field of speech and audio signal processing, pre-trained models (PTMs) are commonly available. Pre-trained models (PTMs) offer a collection of initial weights and biases that may be adjusted for a particular task, which makes them a popular starting point for ML model development. State-of-the-art performance in speech recognition, natural language processing, and other applications has been shown using pre-trained model representations. Embeddings obtained from these models are used as inputs for learning algorithms that are used for a variety of downstream tasks. This study compares pretrained models to show how they perform in Automatic Speech Recognition (ASR). The literature research indicates that self-supervised models based on Wav2Vec2.0 and fully supervised models such as Whisper are the basic paradigms and approaches for ASR currently. This study evaluated and compared these strategies in order to check how well they perform across a wide range of test scenarios. This survey aims to serve as a practical manual for understanding, using, and generating PTMs for different NLP tasks.

Keywords

PTMs, ASR, Wav2vec2, Whisper, Speech Recognition, Natural Language Processing.

1. Introduction

Recent developments in ASR have brought about unique end-to-end architectures(Amodei et al., 2016) that have shown to be accurate enough under such challenging circumstances. End-to-end models' fundamental concept is

to directly translate the input speech signal to character sequences, substantially simplifying training, fine-tuning, and inference generation (Chan et al., 2015; Chorowski et al., 2015; Graves & Jaitly, n.d.; L. Lu et al., 2016; Yao et al., 2021). Fully supervised and self-supervised models are the two basic strategies for training end-to-end ASR systems. In order to provide an end-to-end ASR model that is both competitive and lighter, NVIDIA introduced Quartznet (Kriman et al., 2019). The QuartzNet architecture consists of multiple blocks of 1D convolutions connected by residual connections, as illustrated in **Figure 1**. The word error rates (WERs) for the model ranged from 7.7% to 12.5%, depending on the language, when trained and evaluated on the Common Voice corpus (Bermuth et al., 2021). Additionally, the QuartzNet model achieved WERs of 19.2% for French and 18.3% for Spanish multimedia data from the MediaSpeech corpus (Jia Deng et al., 2009).. Citrinet (Majumdar et al., 2021) was recently introduced by NVIDIA researchers as an advancement of QuartzNet. This model integrates sub-word encoding, a squeeze-and-excitation mechanism, and a residual network utilizing 1D time-channel separable convolutions (Hu et al., 2017). The authors reported a word error rate (WER) of 5.6% on the TEDLIUMv2 corpus.

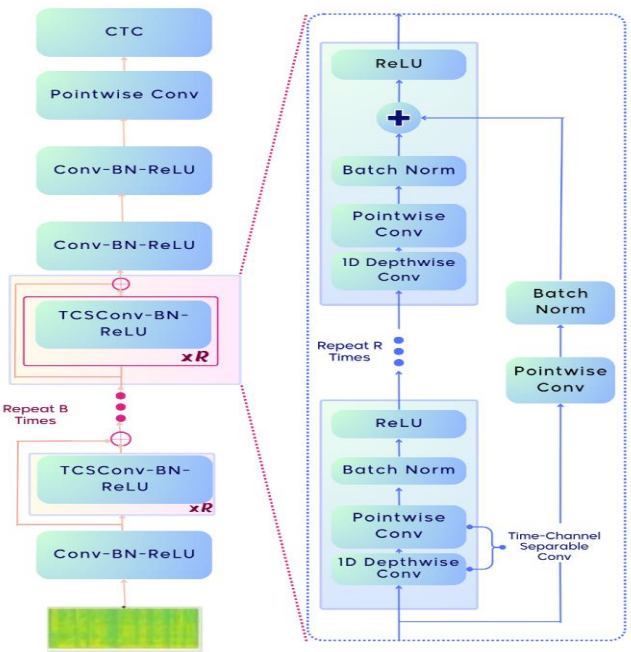

Figure 1. QuartzNet BxR architecture.

In contrast to fully supervised models, recent studies have concentrated on the use of large-scale acoustic models trained using self-supervised learning techniques and a significant amount of unlabeled data. By introducing Wav2Vec2.0 (Baevski et al., 2020), researchers from Meta AI illustrated the capability of models of this type. When compared to benchmark results, Wav2Vec2.0 performed better, especially when ASR for low-resource languages in the Common Voice corpus was taken into account (Pham et al., 2022). In particular, the authors in (Krabbenhöft & Barth, 2022) took into account a Wav2vec2.0 model in combination with their suggested language modelling approach and obtained cutting-edge results in the German Common Voice corpus with a WER of 3.7%. Additionally, Wav2Vec2.0-based models have been evaluated effectively in more challenging acoustic conditions, such as using multimodal Portuguese data from the CORAA database (CERQUEIRA BISPO DOS SANTOS, 1997). Due to these factors, Wav2Vec2.0 has become one of the neural-based models for ASR that is most commonly studied. Self-supervised techniques like Wav2Vec 2.0 are complex because the input sound units during the pre-training phase lack a predefined lexicon. Additionally, sound units vary in length and have unclear segmentation. To address these issues, Meta AI introduced HuBERT (Hsu et al., 2021) as a new method for learning self-supervised speech representations. In many ASR scenarios, the convolutional and transformer networks from Wav2Vec2.0 and HuBERT are combined to achieve state-of-the-art results. The "convolutional augmented transformer" or Conformer was developed by Google researchers to combine the best features of both types of networks into a single neural block (Gulati et al., 2020). In the TEDLIUMv2 corpus, a Conformer network produced a WER of 7.2% (Guo et al., 2020).

Wav2Vec2.0, HuBERT, and Conformers are examples of self-supervised audio encoders that learn superior audio representations. However, because the pre-training was unsupervised, they lacked a suitable decoder to convert these representations into useful outputs. In order to accurately implement models for ASR or audio

classification, a fine-tuning stage is always required. OpenAI researchers have suggested "Whisper" (Radford et al., 2022a) as a potential solution to the aforementioned issue. Whisper is a fully supervised sequence-to-sequence transformer that was trained using up to 680,000 hours of labelled audio from the Internet. On numerous benchmark datasets for ASR, including librispeech, TEDLIUM, and Common Voice, among others, the model has produced state-of-the-art WER results.

The literature research indicates that self-supervised models based on Wav2Vec2.0 and fully supervised models like Whisper are the two basic paradigms and approaches for ASR currently. This study evaluates and compare these strategies in order to check how well they performed robust ASR across a wide range of test scenarios. The remaining portions of the paper are divided as follows; different technological facets of the pretrained models' architectures for ASR are described in section 2. Section 3 discusses the main insights obtained from results. Finally, the conclusions are demonstrated in section 4.

2. Methods

Self-supervised models based on Wav2Vec2.0 and fully supervised models such as Whisper are the two major paradigms and approaches for ASR so far (Vásquez-Correa & Álvarez Muniain, 2023). The present research examined and contrasted these two methods in order to evaluate how well they performed for reliable ASR.

In the upcoming subsections, Different models that play a crucial role in the field of ASR will be explored. Each model brings its unique approach and advancements to improve speech recognition accuracy and performance. In subsection 2.1, the Wav2vec2.0 model, which is utilized with self-supervised learning techniques, is delved into. Subsection 2.2 introduces the Whisper model, a highly efficient ASR system designed for low-resource languages. subsection 2.3 the Whisper-AT model is discussed. The Hidden-Unit BERT (HuBERT) self-supervised speech representation learning strategy is demonstrated in subsection 2.4. In subsection 2.5, the SpeechStew model is explored. Conformer which integrate convolutional neural networks and transformers to efficiently model both local and global dependencies within an audio sequence. are explained in subsection 2.6. In subsection 2.7, unique CNN-RNN-transducer architecture, which called ContextNet is explained. Finally in subsection 2.8 The encoder-decoder RNN is explored.

2.1. Wav2vec2.0

Wav2Vec 2.0 is an end-to-end architecture that employs self-supervised learning. constructed from transformer and convolutional layers as shown in **Figure 2**. The model uses a multi-layer convolutional feature encoder $f: \chi \rightarrow Z$ to transform raw audio waveforms χ into latent speech representations z1,....., and zT .These latent representations supplied the network g: $Z \rightarrow C$, which was transformer-masked. The goals in the self-supervised learning objective are represented by the discrete set of outputs q1,..., qT that are formed when the transformer network first quantizes the continuous representations [(Baevski et al., 2020), (Conneau et al., 2020)]. These quantized representations are then contextualized using the transformer module's attention blocks to produce a collection of discrete contextual representations, symbolized by the letters c1,..., cT. Seven convolutional blocks with 512 channels, kernel widths of {10, 3, 3, 3, 3, 2, 2} and strides of {5, 2, 2, 2, 2, 2, 2} form the feature encoder. The transformer network is composed of 24 blocks, 1024 dimensions, 4096 inner dimensions, and 16 attention heads in total.

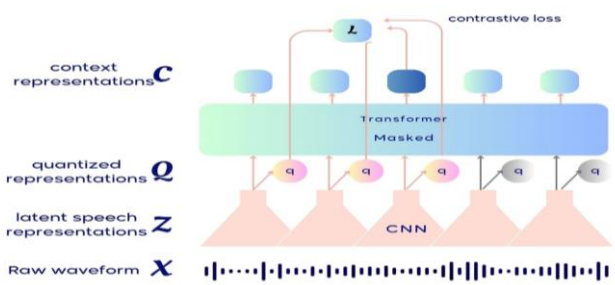

Figure 2. Architecture representation for wav2vec2.0. The raw audio signal is transformed to speech representations that are fed into a network of transformer to output context representations. Figure derived from the one shown in (Baevski et al., 2020).

Wav2vec2 can be used for a variety of speech downstream tasks such as automatic speech recognition ,detection, speaker recognition and language detection. The authors (Baevski et al., 2020) get a WER of 4.8/8.2 on test-clean/other of Librispeech using only 10 minutes of labeled training data, or 48 recordings of 12.5 seconds on average, illustrating the significant potential of pre-training on unlabeled data for speech processing. Wav2vec 2.0 improves upon the previous best result on the clean 100-hour Librispeech while applying 100 times less labeled data.

Authors in (Baevski et al., 2020) perform two models that have different Transformer setups but the same encoder architecture. **Table 1** shows Transformer setups for the two models.

Table 1. Transformer setups for wav2vec2 models.

Size	Number of transformer blocks	Model dimension	Inner dimension	Attention Heads
Base	12	768	3072	8
Large	24	1024	7096	16

A pre-trained Wav2vec2.0 acoustic model based on the Wav2Vec2-XLSR-300M model was taken into consideration by (Vásquez-Correa & Álvarez Muniain, 2023). The model was pre-trained in a self-supervised manner that using 436,000 hours of unlabeled speech data across in 128 languages. Authors in (Vásquez-Correa & Álvarez Muniain, 2023) suggested applying speech recognition and keyword spotting technologies to forensic scenarios, especially in situations involving child exploitation.

2.2. Whisper

OpenAI (Radford et al., 2022a) just released their Whisper ASR system. Whisper is trained in a fully supervised method, as opposed to Wav2vec2.0, using up to 680k hours of labelled speech data from various sources. The encoder-decoder Transformer (Vaswani et al., 2017) that serves as the model's foundation receives 80-channel log-Mel spectrograms as input. This 80-channel log-magnitude Mel spectrogram representation is produced on 25-millisecond windows and a stride of 10 milliseconds after all audio has been re-sampled to 16k Hz. The input representation is processed by the encoder using a tiny stem that consists of 2convolution layers with a filter width of 3 and the GELU activation function (Hendrycks & Gimpel, 2016), where the second convolution layer has a stride of 2. The output of the stem is then supplemented with sinusoidal position embeddings, and then the encoder Transformer blocks are used. The encoder output is processed through a final layer normalization after the transformer applies pre-activation residual blocks (Child et al., 2019). The decoder employs coupled input-output token representations (Press & Wolf, n.d.) and learned position embeddings. The width and number of transformer blocks are the same for the encoder and decoder. The general architecture of Whisper is shown in **Figure 3**.

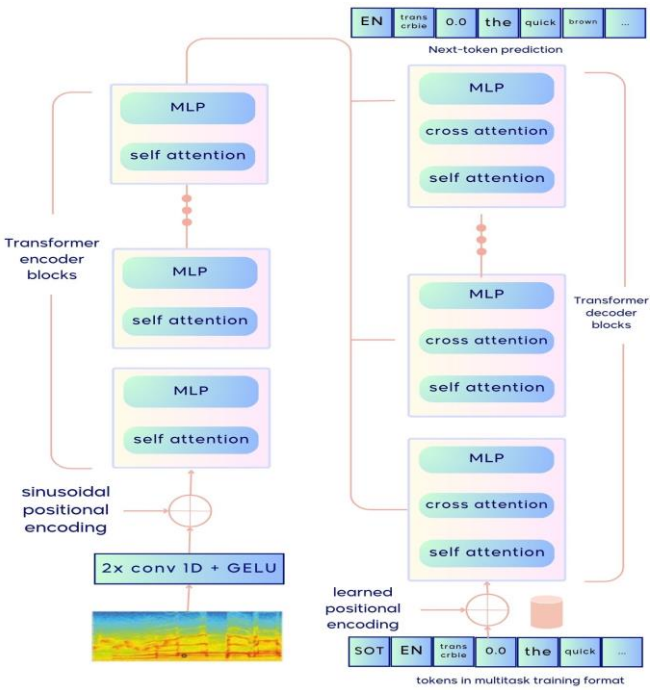

Figure 3. Architectural illustration in whispers. A transformer network encodes the log Mel-spectrograms. The transformer decoder converts encoded representations into character outputs and no-speech tokens.

There are various pre-trained models available with varying attention heads and numbers of layers. **Table 2** shows the whisper model family's architectural details(Radford et al., 2022b).

Table 2. The Whisper model family's architectural details.

Size	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

Whisper has a lot of features, such as Automatic speech recognition, a multi-task model, the ability to perform speech translation and language identification, training on a large dataset of diverse audio, and multilingual speech recognition. However, when it comes to building production systems at scale involving real-time processing of streaming voice data, there are a number of considerations that may make Whisper less suitable than commercially available ASR solutions. Some of its notable limitations include Whisper being slow and expensive, Only Large-v2 being available via API (Tiny, Base, Small, and Medium models are excluded), and limited entity formatting.

Whisper maintains that scaling weakly supervised pretraining has received little attention in speech recognition research up to this point. The researchers (Radford et al., 2022b) indicated how training on a large and diverse supervised dataset, while emphasizing zero-shot transfer, can significantly improve the robustness of a speech recognition system. They achieve results without relying on self-supervision or self-training techniques which have been a mainstay of recent large-scale speech recognition work.

2.3. Whisper-AT

(Gong et al., 2023) first demonstrate that while Whisper is relatively resistant to background sounds (such as music), its audio representation is really not noise-invariant but rather strongly correlated to non-speech sounds, suggesting that Whisper detects speech with depending on the noise type. By freezing the Whisper backbone and training a lightweight audio tagging model on top , they were able to create the combined audio tagging and speech recognition model ,Whisper-AT. Whisper-AT can identify audio events alongside spoken text with only a 1% increase in computing cost.

The authors used Whisper to create a unified model for ASR and Audio Tagging that concurrently recognizes spoken text and background noises (such as music, horns, etc.), which is extremely desirable in applications like video transcription, voice assistants, and hearing aids. Whisper is the best choice for the foundation of such a unified model because: it is resistant to background noise; and its intermediate representations encode detailed general audio event information, providing a strong foundation for audio tagging. To be able to forecast a sound class, they must train a model on top of the Whisper intermediate representations because the original Whisper does not output sound labels.

In order to preserve the Whisper ASR capability and enable the generation of text and audio labels in a single forward pass, they intentionally do not change the original weights of the Whisper model but rather add new audio tagging layers on top of it. This combined ASR and Audio Tagging model is known as Whisper-AT. As shown in **Figure 4.** , (Gong et al., 2023) suggest the following effective design: (1) They add a mean pooling layer to each Whisper representation to reduce the time sequence length n from 500 to 25; (2) They optionally add a linear projection layer to reduce d from 1280 to 512 before audio tagging Transformers (denoted by TL-Tr512); and (3) For WA-Tr, they first conduct weighted averaging and then apply a temporal Transformer, for TL-Tr, they employ just one temporal Transformer for all layers. Thus, just one temporal Transformer is required for both WA-Tr and TL-Tr.

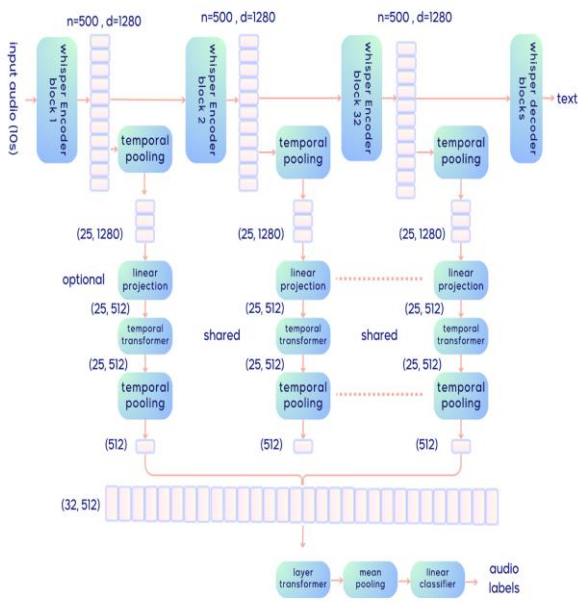

Figure 4. Model of the proposed time- and layer-wise Transformer. Figure derived from the one shown in (Gong et al., 2023).

2.4. HuBERT

Self-supervised methods for learning speech representations are complicated by three particular issues: Each input utterance has several sound units, there is no lexicon of input sound units during the pre-training process, and sound units have varied lengths without being explicitly segmented. (Hsu et al., 2021)suggest the Hidden-Unit BERT (HuBERT) strategy for self-supervised speech representation learning to solve these three issues. This approach makes use of an offline clustering step to generate aligned target labels for Bidirectional Encoder Representations from Transformers (BERT)-like prediction loss. Applying the prediction loss just to the masked regions encourages the model to develop a unified acoustic and linguistic model for continuous inputs, which is a crucial component of the methodology. Instead of depending on the intrinsic quality of the assigned cluster labels, HuBERT primarily depends on the reliability of the unsupervised clustering stage. With 10min, 1h, 10h, 100h, and 960h fine-tuning subsets, the HuBERT model either matches or surpasses the state-of-the-art performance of Wav2Vec 2.0 on the Librispeech (960h) and Libri-light (60kh) benchmarks. This is done by starting with a basic k-means teacher of 100 clusters and utilizing two iterations of clustering. On the more difficult dev-other and test-other evaluation subsets, HuBERT demonstrates up to 19% and 13% relative reductions in WER when using a 1B parameter model.

(Hsu et al., 2021) show organized findings for three HuBERT-pretrained model sizes: BASE (90M parameters), LARGE (300M), and X-LARGE (1B). When pre-trained on the Libri-Light 60k hours, the X-LARGE model exhibits up to a 19% and 13% relative WER improvement over LARGE models on the dev-other and test-other evaluation subsets, respectively.

The authors use the wav2vec 2.0 architecture, which includes a code embedding layer, a projection layer, a BERT encoder (Devlin et al., 2018) , and a convolutional waveform encoder. Similar to the scale of the Conformer XXL model in (Y. Zhang et al., 2020), the X-LARGE architecture increases the model size to around 1 billion parameters. The waveform encoder, which consists of seven 512-channel layers with strides [5,2,2,2,2,2] and kernel widths [10,3,3,3,3,2,2], is the same for all three configurations. The BERT encoder is made up of numerous identical transformer blocks, whose properties are listed in **Table 3** along with those of the following projection layer.

Table 3. HuBERT model architectural summaries for BASE, LARGE, and X-LARGE.

Size	Layers	Width	Heads	Parameters
Base	12	768	8	95M
Large	24	1024	16	317M
X-Large	48	1280	16	964M

The authors in (Hsu et al., 2021) suggest using acoustic unit discovery models to generate frame-level targets illustrated in **Figure 5**. Let X stand for a speech utterance with the form X = [x1, , xT] of T frames. The notation for found hidden units is h(X) = Z = [z1, , zT], where zt∈[C] denotes a categorical variable of the C-class, and h denotes a clustering model, such as k-means.

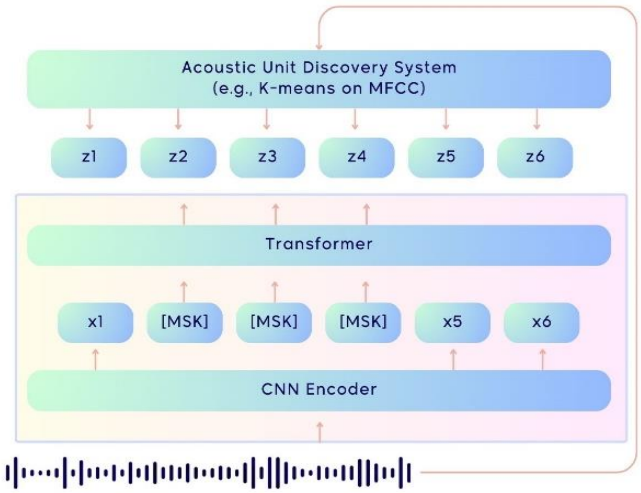

Figure 5. By using one or more iterations of k-means clustering to create the masked frames (y2, y3, and y4), the HuBERT technique predicts hidden cluster assignments.

2.5. SpeechStew

(Chan et al., 2021) introduce SpeechStew, a speech recognition model trained on several publicly accessible speech recognition datasets, including AMI, Broadcast News, Common Voice,LibriSpeech, Switchboard/Fisher, Tedlium, and Wall Street Journal. Without performing any additional re-weighting or re-balancing, SpeechStew simply combines all of these datasets. SpeechStew gets SoTA or close to SoTA results on a number of tasks,without using an external language model. Their results greatly outperform earlier work with powerful external language models, with 9.0% WER on AMI-IHM, 4.7% WER on Switchboard, 8.3% WER on CallHome, and 1.3% on WSJ. Additionally, (Chan et al., 2021) show that SpeechStew picks up strong transfer learning representations. On the CHiME-6 noisy low resource speech dataset, they fine-tuned SpeechStew. Without a language model, they achieve 38.9% WER, compared to a language model's 38.6% WER to a strong HMM baseline.

The Conformer (Gulati et al., 2020) RNN-T (Graves, 2012a) architecture is used by SpeechStew. They test the 1B parameter configuration (Y. Zhang et al., 2020) as well as the 100M parameter (Gulati et al., 2020) . (Chan et al., 2021) discover that the 1B parameter model requires wav2vec pre-training (Baevski et al., 2020). The learning rate schedule and other default hyperparameters from earlier work are used. They don't use an external language model.

2.6. Conformer

Recently, models based on transformer and convolution neural networks (CNNs) have outperformed recurrent neural networks (RNNs) in automatic speech recognition (ASR), demonstrating encouraging results. CNNs effectively use local features, while Transformer models are good at capturing content-based global interactions. (Gulati et al., 2020) investigate how to combine transformers and convolution neural networks to model both local and global dependencies of an audio sequence in a parameter-efficient manner, achieving the best of both worlds. In this regard, they suggest Conformer, a convolution-augmented voice recognition transformer. Conformer performs noticeably better than the earlier Transformer and CNN-based models, achieving cutting-edge accuracy. The Model obtains a WER of 2.1%/4.3% on the widely used LibriSpeech benchmark without the use of a language model and 1.9%/3.9% with the use of an external language model. On test/testother. Additionally, The authors observe 2.7%/6.3% competitive performance using a tiny model with only 10M parameters.

The authors in (Gulati et al., 2020) investigate how to naturally include self-attention and convolutions in ASR models. In order to be parameter efficient, they hypothesis that both global and local interactions are crucial. To do this, the Authors suggest a unique self-attention and convolutional neural network architecture that combines the best of both worlds: self-attention learns the global interaction, while convolutions effectively capture the relative-offset-based local correlations. They present a novel joining of self-attention and convolution, sandwiched between a pair of feed-forward modules (Y. Lu et al., 2019; Wu et al., 2020) , as shown in **Figure 6**.

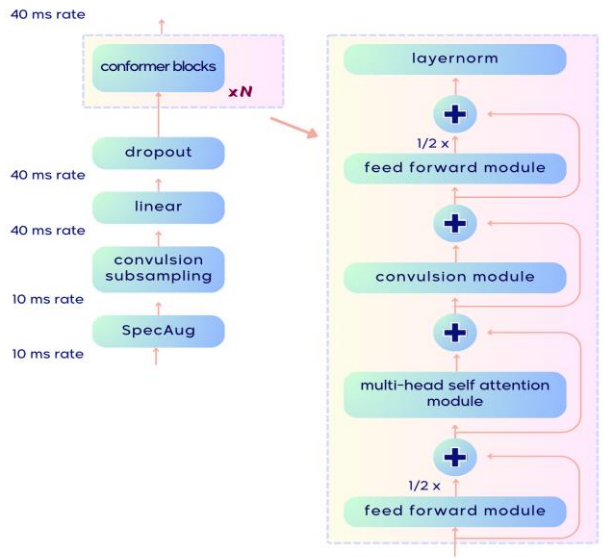

Figure 6. Architecture for the conformer encoder model. The conformer consists of two feed-forward layers that resemble macarons, with halfstep residual connections separating the convolution and multi-headed self attention modules. The post layer norm comes after that.

By comparing various combinations of network depth, model dimensions, and attention head count. The authors (Gulati et al., 2020) discover three models small, medium, and large with 10M, 30M, and 118M parameters, respectively. They then select the model with the greatest performance given the model parameter size restrictions. All of these models employ a single-LSTM-layer decoder. Their architecture hyper-parameters are described in **Table 4**.

Table 4. Model hyper-parameters for Conformer S, M, and L models were discovered by sifting through various combinations and picking the models with the highest performance within the parameter constraints.

Size	Encoder Layers	Encoder Dim	Heads	Parameters	Conv Kernel Size	Decoder Layers	Decoder Dim
Small	16	144	4	10.3M	32	1	320
Medium	16	256	4	30.7M	32	1	640
Large	17	512	8	118.8M	32	1	640

Without a language model, the medium model outperforms the most well-known Transformer, LSTM-based model, or a similar-sized convolution model with competitive results of 2.3/5.0 on test/testother. The model obtains the lowest WER among all the current models when the language model is included. The model obtains a new state-of-the-art performance at 1.9%/3.9% for test/testother and shows improved accuracy with less parameters than prior work on the LibriSpeech dataset.

2.7. ContextNet

The performance of RNN/transformer based models still outperforms that of convolutional neural networks (CNN), despite the latter's promising end-to-end speech recognition results. In this research, (Han et al., n.d.)investigate a unique CNN-RNN-transducer architecture, which they refer to as ContextNet, to bridge this gap and go beyond it ContextNet has a fully convolutional encoder that adds squeeze-and-excitation modules to convolution layers to incorporate global context information. Additionally, they provide a straightforward scaling technique that grows ContextNet's widths and achieves a fair balance between computation and accuracy.

The authors in (Han et al., n.d.) show that ContextNet obtains a WER of 2.1%/4.6% on the clean/noisy LibriSpeech test sets without external language model (LM), 1.9%/4.1% with LM, and 2.9%/7.0% with only 10M parameters. This contrasts with the best model that was previously reported, which had 20M parameters and an LM of 2.0%/4.6% and 3.9%/11.3%, respectively. A much bigger internal dataset is also used to confirm the superiority of the proposed ContextNet model. This paper's primary contributions are (1) an enhanced CNN architecture with global context for ASR, and (2) a progressive downsampling and model scaling approach to achieve greater accuracy and model size trade-off.

The RNN-Transducer architecture (Graves, 2012b; He et al., 2018; Rao et al., 2018a) acts as the network's core of the ContextNet architecture. Three components make up the network: an audio encoder for the input utterance, a

label encoder for the input label, and a joint network to combine them and perform decoding.

ContextNet appears to outperform previously published systems, according to the data. The medium model, ContextNet(M), only has 31M parameters and achieves comparable WER in comparison to considerably bigger systems (Kingma & Ba, 2014; Q. Zhang et al., 2020) . ContextNet(L), The large model, performs better than the prior SOTA by 13% comparatively on test-clean and 18% relatively on test-other.

2.8. End-to-end ASR with RNN-Transducer (RNN-T)

(Rao et al., 2018b) proposed an encoder-decoder RNN model. The proposed method utilizes an encoder network composed of multiple blocks of LSTM layers, pre-trained with CTC to produce phonemes, graphemes, and words as outputs. Additionally, a 1D-CNN reduces the length T of the time sequence by a factor of 3 through specific kernel strides and sizes.The decoder network is an RNN-T model trained alongside an LSTM language model that also predicts words. The network's target is the next label in the sequence, which is used in the cross-entropy loss to optimize the model.Regarding feature extraction, 80-dimensional mel-scale features are calculated every 10 milliseconds and stacked every 30 milliseconds to form a single 240-dimensional acoustic feature vector.

The method is trained on 22 million hand-transcribed audio recordings sourced from Google US English voice traffic, totaling 18,000 hours of training data. This includes both voice search and voice dictation utterances. The language model was pre-trained using text sentences from the dataset. The method was evaluated with various configurations and achieves a 5.2% WER on this extensive dataset when the encoder has 12 layers with 700 hidden units, and the decoder consists of 2 layers with 1000 hidden units each.

3. Discussion

The authors in (Radford et al., 2022b) tested and showed how Whisper compares against other models. The authors don't just want to create a model that outperforms all State-of-the-Art (SOTA) models on a dataset x. Their goal is also to create a model that comfortably generalizes on similar datasets. **Figure 7**. shows comparing effective robustness in great detail across several datasets.

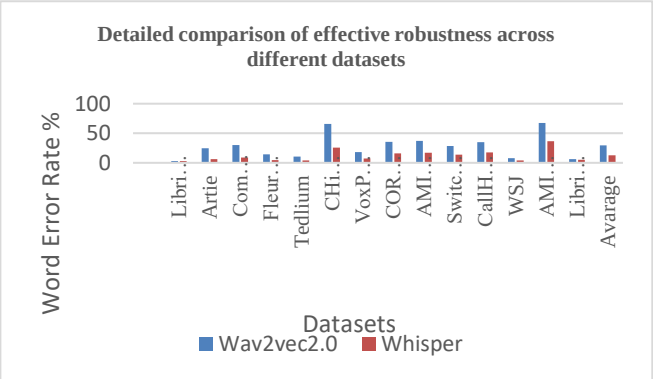

Figure 7. Detailed comparison of effective robustness across different datasets.

The paper compares Whisper against a variant of the Wav2vec2.0 model, in terms of WER — less is better. The task here is transcription. Wav2vec2.0 is specifically trained and fine-tuned on the LibriSpeech dataset, while Whisper is not. On the LibriSpeech dataset, the two models are equal. However, Whisper outperforms Wav2vec2.0 on every other dataset by a large margin.

Similarly, the authors compared Whisper against other SOTA models on the translation task. **Table 5** shows the results. Except for the low resource settings, Whisper outperforms all other models. On average, Whisper achieves a higher BLEU score. Remember that Whisper is utilized in a zero-shot configuration but the other models have been fully trained using their respective datasets.

Table 5. Comparing Whisper models to other SOTA models for audio-to-English translation using the BLEU score.

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	25.2
Zero-Shot Whisper	36.2	32.6	25.2	29.1

4. Conclusion

This paper introduces comparative study on end to end speech using pretrained models such as Whisper and Wav2vec2.0. Modern ASR systems based on Wav2Vec2.0 and Whisper were taken into consideration for the addressed application. A sizable collection of open benchmark corpora was used to show the effectiveness of both methods. As a result, the outcomes can be applied to additional ASR areas.

Wav2vec2 can be applied to various speech downstream tasks such as automatic speech recognition, detection, speaker recognition and language detection. Whisper has a lot of features, such as Automatic speech recognition, a multi-task model, the ability to perform speech translation and language identification, training on a large dataset of diverse audio, and multilingual speech recognition. Whisper AT model can be used for ASR and Audio Tagging. HuBERT used to solve three particular issues: Each input utterance has several sound units, there is no lexicon of input sound units during the pre-training process, and sound units have varied lengths without being explicitly segmented. The Speech Stew model combines several publicly accessible speech recognition datasets without performing any additional re-weighting or re-balancing for its.

Conformer model can be used which it is a combination of self-attention and convolution modules to achieve the best of the two approaches.

The comparison of the pretrained models and Whisper models showed that the second one was typically the most accurate system.

References

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, K., Du, N., Elsen, E., … Zhu, Z. (2016). Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin.

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. http://arxiv.org/abs/2006.11477

Bermuth, D., Poeppel, A., & Reif, W. (2021). Scribosermo: Fast Speech-to-Text models for German and other Languages. http://arxiv.org/abs/2110.07982

CERQUEIRA BISPO DOS SANTOS, S. (1997). RECONHECIMENTO DE VOZ CONTÍNUA PARA O PORTUGUÊS UTILIZANDO MODELOS DE MARKOV ESCONDIDOS [PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO]. https://doi.org/10.17771/PUCRio.acad.8372

Chan, W., Jaitly, N., Le, Q. V., & Vinyals, O. (2015). Listen, Attend and Spell. http://arxiv.org/abs/1508.01211

Chan, W., Park, D., Lee, C., Zhang, Y., Le, Q., & Norouzi, M. (2021). SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network. http://arxiv.org/abs/2104.02133

Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). Generating Long Sequences with Sparse Transformers. http://arxiv.org/abs/1904.10509

Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-Based Models for Speech Recognition. http://arxiv.org/abs/1506.07503

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2020). Unsupervised Cross-lingual Representation Learning for Speech Recognition. http://arxiv.org/abs/2006.13979

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. http://arxiv.org/abs/1810.04805

Gong, Y., Khurana, S., Karlinsky, L., & Glass, J. (2023). Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers. http://arxiv.org/abs/2307.03183

Graves, A. (2012a). Sequence Transduction with Recurrent Neural Networks. http://arxiv.org/abs/1211.3711

Graves, A. (2012b). Sequence Transduction with Recurrent Neural Networks. http://arxiv.org/abs/1211.3711

Graves, A., & Jaitly, N. (n.d.). Towards End-to-End Speech Recognition with Recurrent Neural Networks.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer:

Convolution-augmented Transformer for Speech Recognition. http://arxiv.org/abs/2005.08100

Guo, P., Boyer, F., Chang, X., Hayashi, T., Higuchi, Y., Inaguma, H., Kamo, N., Li, C., Garcia-Romero, D., Shi, J., Shi, J., Watanabe, S., Wei, K., Zhang, W., & Zhang, Y. (2020). Recent Developments on ESPnet Toolkit Boosted by Conformer. http://arxiv.org/abs/2010.13956

Han, W., Zhang, Z., Zhang, Y., Yu, J., Chiu, C.-C., Qin, J., Gulati, A., Pang, R., & Wu, Y. (n.d.). ContextNet: Improving Convolutional Neural Networks for Automatic Speech Recognition with Global Context.

He, Y., Sainath, T. N., Prabhavalkar, R., McGraw, I., Alvarez, R., Zhao, D., Rybach, D., Kannan, A., Wu, Y., Pang, R., Liang, Q., Bhatia, D., Shangguan, Y., Li, B., Pundak, G., Sim, K. C., Bagby, T., Chang, S., Rao, K., & Gruenstein, A. (2018). Streaming End-to-end Speech Recognition For Mobile Devices. http://arxiv.org/abs/1811.06621

Hendrycks, D., & Gimpel, K. (2016). Gaussian Error Linear Units (GELUs). http://arxiv.org/abs/1606.08415

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. http://arxiv.org/abs/2106.07447

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2017). Squeeze-and-Excitation Networks. http://arxiv.org/abs/1709.01507

Jia Deng, Wei Dong, Socher, R., Li-Jia Li, Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. 248–255. https://doi.org/10.1109/cvprw.2009.5206848

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. http://arxiv.org/abs/1412.6980

Krabbenhöft, H. N., & Barth, E. (2022). TEVR: Improving Speech Recognition by Token Entropy Variance Reduction. http://arxiv.org/abs/2206.12693

Kriman, S., Beliaev, S., Ginsburg, B., Huang, J., Kuchaiev, O., Lavrukhin, V., Leary, R., Li, J., & Zhang, Y. (2019). QuartzNet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions. http://arxiv.org/abs/1910.10261

Lu, L., Zhang, X., & Renais, S. (2016). On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2016-May, 5060–5064. https://doi.org/10.1109/ICASSP.2016.7472641

Lu, Y., Li, Z., He, D., Sun, Z., Dong, B., Qin, T., Wang, L., & Liu, T.-Y. (2019). Understanding and Improving Transformer From a Multi-Particle Dynamic System Point of View. http://arxiv.org/abs/1906.02762

Majumdar, S., Balam, J., Hrinchuk, O., Lavrukhin, V., Noroozi, V., & Ginsburg, B. (2021). Citrinet: Closing the Gap between Non-Autoregressive and Autoregressive End-to-End Models for Automatic Speech Recognition. http://arxiv.org/abs/2104.01721

Pham, N.-Q., Waibel, A., & Niehues, J. (2022). Adaptive multilingual speech recognition with pretrained models. http://arxiv.org/abs/2205.12304

Press, O., & Wolf, L. (n.d.). Using the Output Embedding to Improve Language Models. In the Association for Computational Linguistics (Vol. 2). http://statmt.org/wmt15/translation-task.html

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022a). Robust speech recognition via large-scale weak supervision. ArXiv Preprint ArXiv:2212.04356.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022b). Robust Speech Recognition via Large-Scale Weak Supervision. http://arxiv.org/abs/2212.04356

Rao, K., Sak, H., & Prabhavalkar, R. (2018a). Exploring Architectures, Data and Units For Streaming End-to-End Speech Recognition with RNN-Transducer. http://arxiv.org/abs/1801.00841

Rao, K., Sak, H., & Prabhavalkar, R. (2018b). Exploring Architectures, Data and Units For Streaming End-to-End Speech Recognition with RNN-Transducer. http://arxiv.org/abs/1801.00841

Vásquez-Correa, J. C., & Álvarez Muniain, A. (2023). Novel Speech Recognition Systems Applied to Forensics within Child Exploitation: Wav2vec2.0 vs. Whisper. Sensors, 23(4). https://doi.org/10.3390/s23041843

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L.,

Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. http://arxiv.org/abs/1706.03762

Wu, Z., Liu, Z., Lin, J., Lin, Y., & Han, S. (2020). Lite Transformer with Long-Short Range Attention. http://arxiv.org/abs/2004.11886

Yao, Z., Wu, D., Wang, X., Zhang, B., Yu, F., Yang, C., Peng, Z., Chen, X., Xie, L., & Lei, X. (2021). WeNet: Production oriented Streaming and Non-streaming End-to-End Speech Recognition Toolkit. http://arxiv.org/abs/2102.01547

Zhang, Q., Lu, H., Sak, H., Tripathi, A., McDermott, E., Koo, S., & Kumar, S. (2020). Transformer Transducer: A Streamable Speech Recognition Model with Transformer Encoders and RNN-T Loss. http://arxiv.org/abs/2002.02562

Zhang, Y., Qin, J., Park, D. S., Han, W., Chiu, C.-C., Pang, R., Le, Q. V., & Wu, Y. (2020). Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition. http://arxiv.org/abs/2010.10504