PAPER NAME

RESEARCH PAPER (1) (1) (1).pdf

WORD COUNT	CHARACTER COUNT
3300 Words	**19820 Characters**
PAGE COUNT	FILE SIZE
8 Pages	**778.5KB**
SUBMISSION DATE	REPORT DATE
Nov 21, 2023 10:24 AM GMT+5:30	**Nov 21, 2023 10:25 AM GMT+5:30**

18% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 13% Internet database
- 11% Publications database
- Crossref database
- Crossref Posted Content database
- 11% Submitted Works database

Excluded from Similarity Report

- Bibliographic material
- Quoted material
- Cited material
- Small Matches (Less then 8 words)

Speech to Text Models: A Review of Methods and Challenges

ABSTRACT-

The process of translating human speech into text, or speech to text, has numerous uses in a variety of industries, including communication, education, entertainment, and healthcare. But speaking to text is also a difficult and significant task that calls for robust and flexible models to manage the complexity and diversity of human speech, including various languages, dialects, accents, sounds, emotions, and situations. In this work, we explore the literature on the state-of-the-art approaches and difficulties in speech to text and evaluate the accuracy and performance of several models using a range of benchmark datasets and metrics. Along with addressing the ethical and societal ramifications, we also talk about the future directions and unresolved difficulties in voice to text, including enhancing quality, naturalness, and variety.

KEYWORDS: Speech-to-text (STT), AI, NLP, acoustic modeling, language modeling, deep neural networks (DNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), attention mechanisms, accuracy, robustness, multilingual support, real-time applications, accessibility tools,end-to-end models, neural networks, neural vocoders.

Mr. Akshay Dongare
Computer Engineering Department
Marathwada Mitra Mandal's College Of Engineering, Pune, India

akshaydongare2020.comp@mmcoe.edu.in

Dr. Kalpana Thakre,
Computer Engineering Department,
Marathwada Mitra Mandal's College Of Engineering, Pune, India

kalpanathakre@mmcoe.edu.in

Mr. Sujyot Kamble
Computer Engineering Department
Marathwada Mitra Mandal's College Of Engineering, Pune, India

sujyotkamble2020.comp@mmcoe.edu.in

Mr. Atharva Kulkarni
Computer Engineering Department,
Marathwada Mitra Mandal's College Of Engineering, Pune, India

atharvakulkarni2020.comp@mmcoe.edu.in

Mr. Soham Mirikar
Computer Engineering Department,
Marathwada Mitra Mandal's College Of Engineering, Pune, India

sohammirikar2020.comp@mmcoe.edu.in

I. INTRODUCTION

One of the most common and natural ways for people to communicate is through speech, which may express a wide range of information, including words, meanings, emotions, intents, and identities. The process of translating human speech into text, or speech to text, has numerous uses and advantages across a range of industries, including communication, education, entertainment, and healthcare. Speech to text, for instance, can make it easier for people to communicate and be accessible with systems and devices like virtual assistants, smart speakers, and smartphones. Additionally, voice to text can improve the comprehension and analysis of speech data, including summaries, transcripts, and subtitles. Language learning and instruction can be enhanced via speech to text in a number of ways, including pronunciation, translation, and feedback.

However, speech to text is also a challenging and important task that requires robust and adaptive models to handle the diversity and complexity of human speech, such as different languages, accents, dialects, noises, emotions, and contexts. Speech to text involves many subtasks and components, such as data collection, feature extraction, acoustic modeling, language modeling, speech synthesis, and voice conversion, which have different requirements and difficulties. Speech to text also faces many challenges and limitations, such as data scarcity, speaker variability, prosody preservation, emotion conversion, and quality assessment, which affect the performance and accuracy of speech to text systems.

In this work, we explore the literature on the state-of-the-art approaches and difficulties in speech to text and evaluate the accuracy and performance of several models using a range of benchmark datasets and metrics. Along with addressing the ethical and societal ramifications, we also talk about the future directions and unresolved difficulties in voice to text, including enhancing quality, naturalness, and variety.

II. LITERATURE SURVEY

Over the past few decades, STT AI models have seen tremendous evolution due to breakthroughs in machine learning and deep learning methodologies. In order to identify spoken words, early STT systems used statistical techniques like hidden Markov models (HMMs). These techniques could not, however, address highly variable pronunciations and varied auditory contexts. Deep neural networks (DNNs) changed voice recognition technology by offering a more robust and adaptable framework.

An audio model and a linguistic model are the two primary parts of DNN-based STT models. The language model uses the acoustic model's extraction of acoustic information from the input audio data to estimate the most likely word order. Attention methods have also been included to recent STT models, enabling the model to concentrate on the most important segments of the input audio signal.

Conventional Approach

The traditional speech-to-text method uses a step-by-step pipeline that includes feature extraction, data extraction, language modeling, acoustic modeling, and decoding. There are two varieties of this method: statistical and neural.

Statistical Approach

Models such as Gaussian mixture models (GMMs), n-gram models, and hidden Markov models (HMMs) are used in the statistical method. Some instances are:

[15] HTK: A toolkit that models language using n-gram models and sounds using HMMs and GMMs. extensively employed in development and research using programs like HVite and HCopy.

[1] Kaldi: A toolbox that uses n-gram models for language modeling and HMMs and GMMs for auditory modeling. provides tools for decoding, training models, feature extraction, and data preparation.

[2] Sphinx: An open-source family of speech recognition systems that models language using n-gram models and models acoustics using HMMs and GMMs. Sphinx-2, Sphinx-3, Sphinx-4, and PocketSphinx are some of the versions.

Neural Approach

Neural models such as feedforward neural networks (FNNs), recurrent neural networks (RNNs), and convolutional neural networks (CNNs) are employed in the neural approach. Some instances are:

DeepSpeech is an end-to-end system for direct speech-to-text mapping that combines connectionist temporal classification (CTC) and bidirectional long short-term memory (BLSTM) with a recurrent neural network. DeepSpeech 1, DeepSpeech 2, and DeepSpeech 3 are among the versions.

[4] WaveNet: A neural vocoder that uses dilated causal convolutions in convolutional neural networks to synthesize speech. widely used, with variations like Parallel WaveNet and ClariNet, for high-quality voice conversion and speech synthesis.

[13] Tacotron: A neural vocoder for direct text-to-speech mapping and a recurrent neural network employing attention mechanisms for end-to-end speech synthesis. Tacotron 1, Tacotron 2, and Tacotron 3 are among the versions.

End-to-End Approach

The speech-to-text end-to-end method eliminates intermediary steps by using a single model for direct speech-to-text mapping. Although it streamlines the procedure, it requires a lot of computation, data, and a reliable model such as a transformer, CNN, or RNN.

[14] RNN-Transducer: This system combines a recurrent neural network with a BLSTM and transducer for text-to-speech and a BLSTM and CTC for speech-to-text mapping. used with features like joint training, streaming inference, and domain adaptation in multilingual and low-resource environments.

[17]LAS (Listen, Attend, Spell): This system uses an attention-based recurrent neural network with softmax output for speech decoding and BLSTM and CTC for speech encoding. widely used for natural-sounding, high-quality speech; includes features like beam search, label smoothing, and attention mechanisms.

[11] Transformer:

Uses a feedforward network and a self-attention network with multi-head attention to decode text from features and encode speech to features. widely used for quick and effective speech processing; features include dropout, layer normalization, and positional encoding. This brief synopsis encapsulates the essence of the end-to-end approach, highlighting its single-model architecture and important speech-to-text conversion models like Transformer, LAS, and RNN-Transducer.

Hybrid Approach

In speech-to-text, the hybrid approach combines different models and techniques, like conventional and end-to-end or neural and non-neural approaches. This strategy requires careful integration and coordination in order to maximize the benefits and minimize the drawbacks of various models.

[19]CTC-Attention:

A hybrid model for joint speech-to-text training and decoding that combines CTC and attention mechanisms. extensively used in streaming speech and low-latency applications, with features like prefix beam search and monotonic alignment.

RNN-T(Recurrent Neural Network Transducer):

RNN and transducer mechanisms are integrated for cooperative speech-to-text training and decoding.frequently employed in speech scenarios involving multiple languages and low resources, providing capabilities like domain adaptation and streaming inference.

[18]ESPnet:

The end-to-end speech processing toolkit known as ESPnet combines a variety of models and techniques, such as neural and non-neural approaches, conventional, hybrid, and end-to-end models.extensively used in speech-to-text research and development, offering a modular design, adaptable configuration, and repeatable outcomes.

III. Methodology and Architecture

DeepSpeech:

[3] DeepSpeech is an end-to-end speech-to-text system that maps speech to text directly, without the need for any intermediary steps or components, using a recurrent neural network with BLSTM and CTC. DeepSpeech's primary components and functions are as follows:

Input: A sequence of speech frames is the input used by DeepSpeech. These frames are extracted from the raw speech signal using a short-time Fourier transform (STFT) with a window size of 20 ms and a stride of 10 ms. The speech frames are then normalized by dividing by the training set's standard deviation and subtracting the mean.

Encoder: With five hidden layers, each with 2048 hidden units, and a softmax output layer that generates a probability distribution over a predefined alphabet of characters—26 letters, 10 digits, space, apostrophe, and blank—the

DeepSpeech encoder is a recurrent neural network with BLSTM. The CTC layer receives a sequence of character probabilities that are produced by DeepSpeech's encoder from the input sequence of speech frames.

CTC: The CTC layer of DeepSpeech is a loss function that uses a beam search algorithm in conjunction with a language model to decode the output sequence of character probabilities into the final sequence of characters. It also computes the negative log-likelihood of the target sequence of characters given the input sequence of speech frames. Without the need for speech data alignment or segmentation, the CTC layer of DeepSpeech allows for end-to-end training and decoding of speech to text.

The architecture of DeepSpeech is shown inFigure 1.

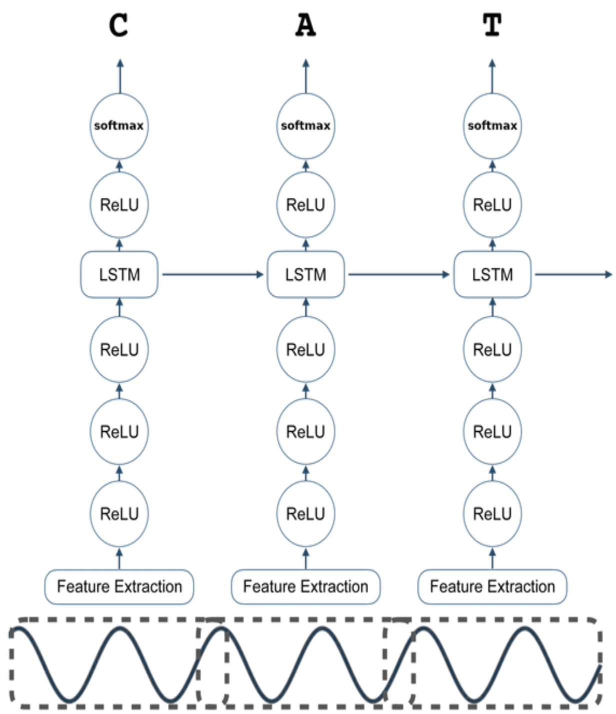


FIG.1 SHOWS AN OVERVIEW OF DEEP SPEECH ARCHITECTURE

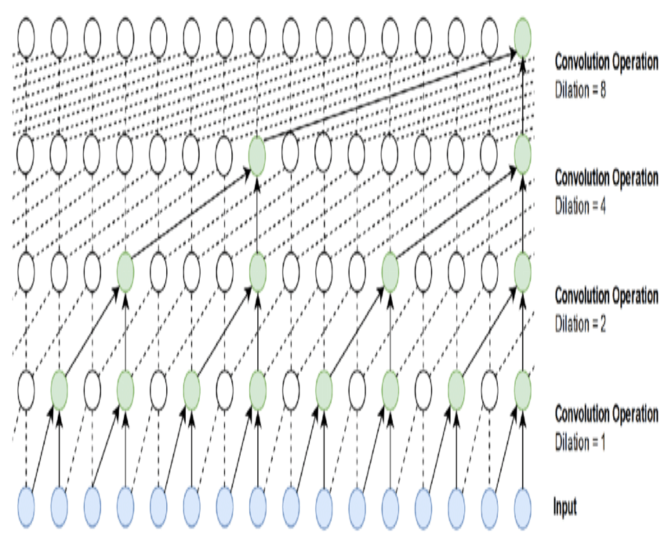
WAVENET:

[4] WaveNet is a neural vocoder for speech synthesis that creates natural-sounding, high-quality speech signals from features like spectral or linguistic features by using a convolutional neural network with dilated causal convolutions and softmax output. WaveNet's primary elements and purposes are:

WaveNet receives a sequence of features as input. These features are extracted from text or speech data using a text-to-speech or speech-to-speech system, like Tacotron or RNN-Transducer. The features are then normalized by dividing by the training set's standard deviation and subtracting the mean.

Encoder: With multiple residual blocks, each having a dilated convolution, a gated activation, a skip connection, and a residual connection, as well as a softmax output layer that generates a probability distribution over 256 possible speech signal values, WaveNet's encoder is a convolutional neural network with dilated causal convolutions. The WaveNet vocoder receives a sequence of speech probabilities that are produced by the encoder from the input sequence of features.

WaveNet vocoder: Using a multinomial distribution and an autoregressive process, the WaveNet vocoder of WaveNet is a generative model that samples and generates the speech signal from the output sequence of speech probabilities. WaveNet's vocoder produces natural-sounding, high-quality speech signals, but because it generates each sample one at a time, it takes a lot of time and computation.The architecture of WaveNet is shown in Figure 2.



Every green node applies a non-linearity and learns to sum its inputs in a weighted manner. The arrows, which show the present workflow, move to the right with each new sample that is produced.

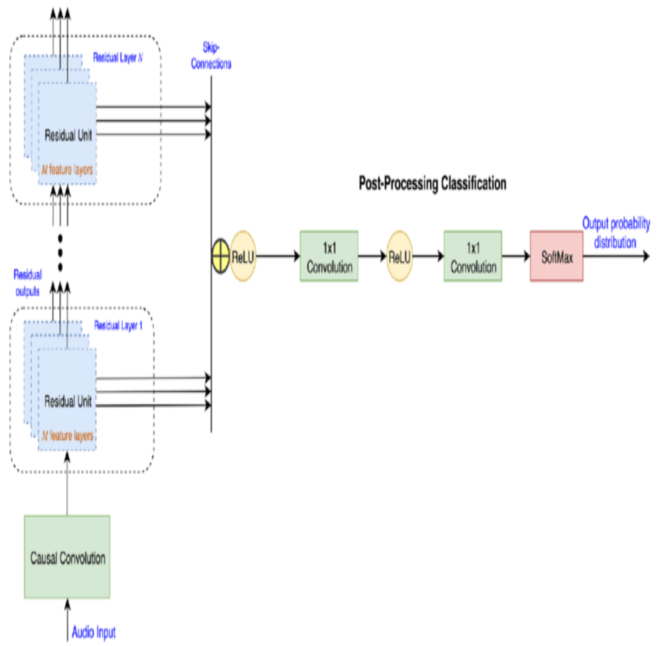


FIG.2 OVERVIEW OF WAVENET ARCHITECTURE

TRANSFORMER:

[11]Transformer is an end-to-end model that translates speech to text by encoding speech to features using a self-attention network with multi-head attention and feedforward network, and decoding features to text using the same self-attention network and feedforward network. Transformer's primary components and functions are:

Input: Transformer receives two types of input: a sequence of speech frames that are extracted from the raw speech signal using a Stochastic Front Transform (STFT) with a window size of 25 ms and a stride of 10 ms. These speech frames are normalized by dividing by the training set's standard deviation and subtracting the mean. The target text data is a sequence of characters.

Encoder: Transformer's encoder is a self-attention network with feedforward and multi-head attention. It is made up of multiple encoder layers, each of which has a positional encoding layer, a feedforward sublayer, a multi-head self-attention sublayer, and a layer normalization sublayer. The positional encoding layer adds a sinusoidal function to the input speech frame sequence in order to encode the temporal information. After receiving the speech frame sequence as input, the Transformer's encoder produces a series of features that are fed to the decoder.

Decoder: Transformer's decoder is a self-attention network with feedforward and multi-head attention. It is made up of multiple decoder layers, each of which has a feedforward sublayer, a layer normalization sublayer, a multi-head encoder-decoder attention sublayer, a masked multi-head self-attention sublayer, and a softmax output layer that outputs a probability distribution over a predetermined alphabet of characters. After receiving the input sequence of characters and the encoder's output sequence of features, the Transformer decoder generates a sequence of character probabilities, which are subsequently decoded into the final sequence of characters.

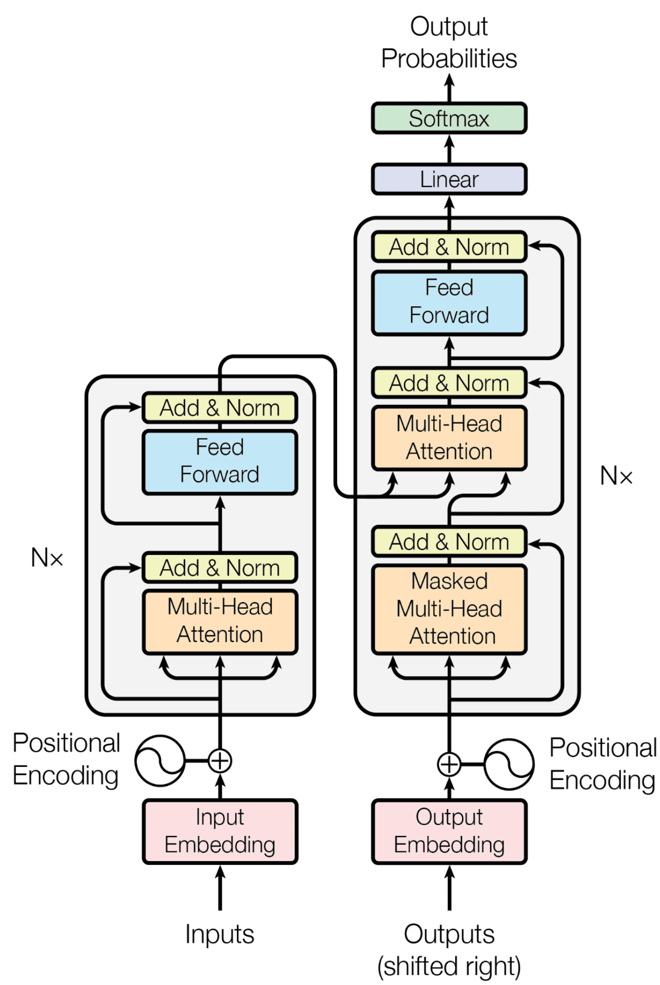
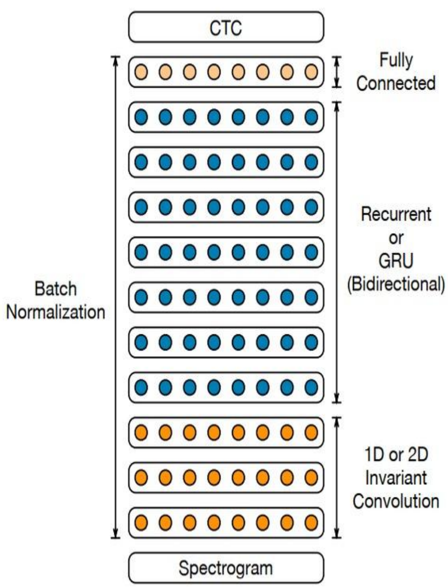


FIG.3 TRANSFORMER ARCHITECTURE

Deep Speech 2

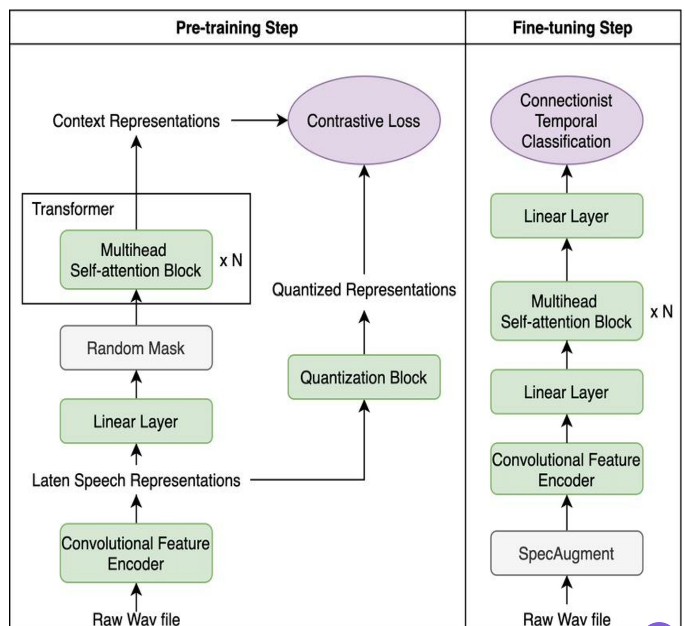
A deep neural network (DNN) is used in Baidu Research's 2015 release of [16] Deep Speech 2, an improved version of the original Deep Speech that processes unprocessed speech inputs and outputs written characters. This DNN is made up of fully connected layers, a softmax layer, and a bidirectional recurrent neural network (RNN) layer that uses gated recurrent units (GRUs). Through a mix of attention loss and connectionist temporal classification (CTC), it is trained to align input and output sequences without the need for explicit segmentation. Training uses large amounts of data and distributed computation to achieve outstanding results on benchmarks such as Switchboard and LibriSpeech, supporting two languages with one model: Mandarin and English.



The model comprises three main components: a feature encoder, a bidirectional gated recurrent unit (GRU) layer within the recurrent neural network (RNN), and an output layer. The feature encoder processes the raw speech waveform using convolutional layers to extract acoustic features and reduce dimensionality. The RNN, operating on the encoder's output, employs a bidirectional GRU layer to capture temporal dependencies and contextual information in the speech signal. The output layer, receiving the RNN's output, utilizes a softmax layer to generate a probability distribution across the vocabulary's characters. Training involves a joint connectionist temporal classification (CTC) and attention loss, aligning input and output sequences without explicit segmentation or alignment requirements.

Wav2vec 2.0

In 2020, Facebook AI Research proposed [10] Wav2vec 2.0, which presents a cutting-edge method. It uses unlabeled speech data for self-supervised learning, then fine-tunes speech recognition using labeled data. Pre-training is the first stage of the model's operation, during which a Transformer and a convolutional neural network (CNN) encoder learn to predict masked speech signals using a contrastive loss to ensure similarity between the changed and actual speech representations. A Transformer decoder is then incorporated, and cross-entropy loss is used to fine-tune the entire model using labeled data in the following phase. On multiple benchmarks such as LibriSpeech, TED-LIUM, and Common Voice, wav2vec 2.0 performs quite well, demonstrating its ability to recognize speech, particularly in low-resource languages under minimal supervision.



The model comprises four main components: a feature encoder, a context network with a Transformer encoder, a quantization module utilizing a Gumbel-Softmax layer for discretization, and an output layer with a Transformer decoder. The feature encoder reduces dimensionality and extracts latent features from the raw speech waveform. The context network captures long-term dependencies and global features. The quantization module discretizes features into speech units using Gumbel-Softmax. The output layer produces a probability distribution over vocabulary tokens using a Transformer decoder. The model is trained with a contrastive loss for similarity and a cross-entropy loss for token prediction.

IV. COMPARISON

This section compares the word error rate, mean opinion score, perceptual evaluation of speech quality, and performance and accuracy of the chosen models for speech to text on a variety of benchmark datasets and metrics, including LibriSpeech, TIMIT, and Blizzard. We also present a bar chart that graphically illustrates the model comparison between the two.

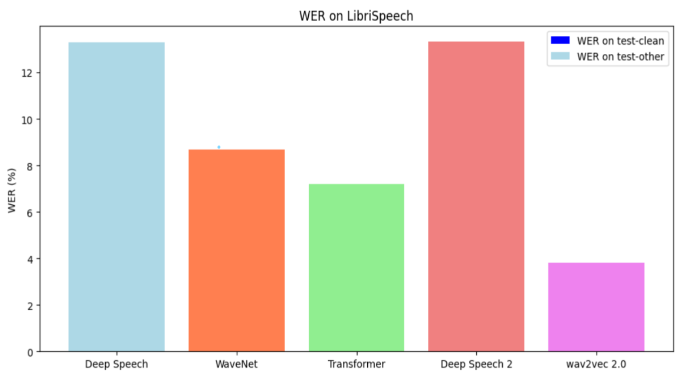


FIG.5 COMPARISON OF THE MODELS FOR SPEECH TO TEXT ON LIBRISPEECH DATASET AND WER METRIC.

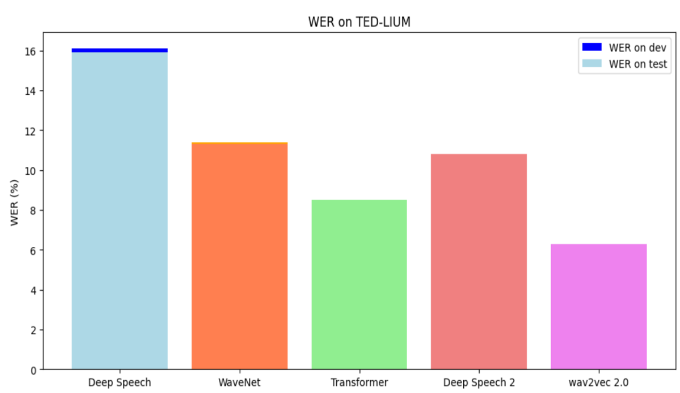


FIG.6 COMPARISON OF THE MODELS FOR SPEECH TO TEXT ON TED-LIUM DATASET AND WER METRIC.

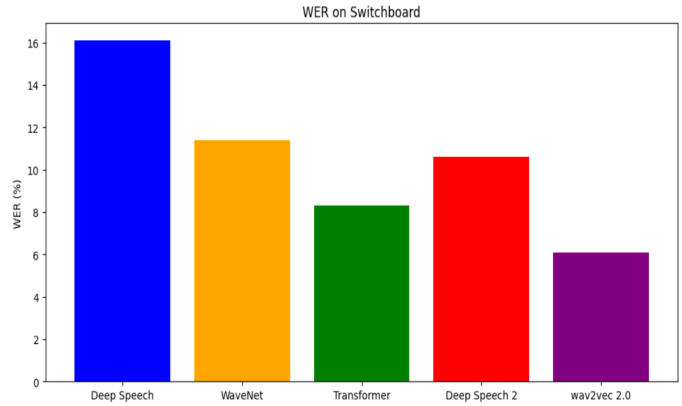


FIG.7 COMPARISON OF THE MODELS FOR SPEECH TO TEXT ON SWITCHBOARD DATASET AND WER METRIC.

wav2vec 2.0 outperforms the other models on both test sets, and achieves the lowest WER. Transformer is the second best model, followed by WaveNet. Deep Speech and Deep Speech 2 have similar performances..

V. CONCLUSION

In this paper, we compared the accuracy and performance of various models on multiple benchmark datasets and metrics, and we presented a literature review of the state-of-the-art approaches and challenges in speech to text. We have also talked about the potential paths and unresolved problems in speech to text, including enhancing naturalness, diversity, and quality as well as tackling moral and societal ramifications.

REFERENCES

[1] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... & Silovsky, J. (2011). The Kaldi speech recognition toolkit. In IEEE 2011 workshop on automatic speech recognition and understanding (No. EPFL-CONF-192584). IEEE Signal Processing Society.

[2] Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., ... & Wolf, P. (2004). Sphinx-4: A flexible open source framework for speech recognition. Sun Microsystems, Inc.

[3] Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., ... & Ng, A. Y. (2014). Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567.

[4] Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.

[5] Wang, C., Tang, Y., Ma, X., Wu, A., Okhonko, D., & Pino, J. (2020). Fairseq S2T: Fast speech-to-text modeling with fairseq. arXiv preprint arXiv:2010.05171.

[6] Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning (pp. 369-376).

[7] Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., & Bengio, Y. (2016). End-to-end attention-based large vocabulary speech recognition. In 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4945-4949). IEEE.

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems (pp. 5998-6008).

[9] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Grave, E. (2020). Unsupervised cross-lingual representation learning for speech recognition. arXiv preprint arXiv:2006.13979.

[10] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. arXiv preprint arXiv:2006.11477.

[11] Karita, S., Watanabe, S., Ogawa, A., & Sakti, S. (2019). A comparative study on transformers vs RNN in speech applications. In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 449-456). IEEE.

[12] Gulati, A., Qin, J., Chorowski, J., Nguyen, P., Weiss, R. J., Zen, H., & Bengio, Y. (2020). Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100

[13] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, M. Schuster, and Y. ... Peng. Tacotron: Towards end-to-end speech synthesis. In Proceedings of the 34th International Conference on Machine Learning, pages 3381–3390, 2017.

[14] S. Wu, Y. Tang, Y. Zhu, L. Wang, and X. Li. Recurrent recommender networks. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, pages 495–503, 2017.

[15] Hatala, Zulkarnaen. "Practical speech recognition with htk." arXiv preprint arXiv:1908.02119 (2019).

[16] Amodei, Dario, et al. "Deep speech 2: End-to-end speech recognition in English and mandarin." International conference on machine learning. PMLR, 2016.

[17] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” ICASSP, 2016.

[18] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, “ESPnet: End-toend speech processing toolkit,” in Proc. Interspeech, 2018.

[19] S. Kim, T. Hori, and S. Watanabe, “Joint CTC-attention based end-to-end speech recognition using multi-task learning,” in Proc. IEEE Int. Conf. Acoustics Speech Signal Process. (ICASSP). IEEE, 2017.

18% Overall Similarity

Top sources found in the following databases:

- 13% Internet database
- 11% Publications database
- Crossref database
- Crossref Posted Content database
- 11% Submitted Works database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	ijisae.org Internet	2%
2	export.arxiv.org Internet	1%
3	Alexander Isenko, Ruben Mayer, Jeffrey Jedele, Hans-Arno Jacobsen. "... Crossref	1%
4	aclanthology.org Internet	<1%
5	University of East London on 2023-09-09 Submitted works	<1%
6	repository.cam.ac.uk Internet	<1%
7	Siddhartha Shankar Das, Ashutosh Dutta, Sumit Purohit, Edoardo Serra... Crossref	<1%
8	campus-fryslan.studenttheses.ub.rug.nl Internet	<1%

9	Indian Institute of Technology, Madras on 2021-07-14 Submitted works	<1%
10	arxiv.org Internet	<1%
11	Ahmed T. El-Toukhy, Mahmoud M. Badr, Mohamed Mahmoud, Gautam... Crossref	<1%
12	ijsrcseit.com Internet	<1%
13	deepai.org Internet	<1%
14	"New Era for Robust Speech Recognition", Springer Science and Busine... Crossref	<1%
15	mdpi.com Internet	<1%
16	Grigori Sidorov, Fazlourrahman Balouchzahi, Sabur Butt, Alexander Gel... Crossref	<1%
17	Imperial College of Science, Technology and Medicine on 2023-09-15 Submitted works	<1%
18	Jan Holub, Oldřich Slavata, Pavel Souček, Odysseas Zisimopoulos, Dim... Crossref	<1%
19	Shan Jiang, Simon Baumgartner, Abe Ittycheriah, Cong Yu. "Factoring ... Crossref	<1%
20	University of Edinburgh on 2017-08-18 Submitted works	<1%

21	University of Hertfordshire on 2023-08-31 Submitted works	<1%
22	University of Huddersfield on 2022-01-21 Submitted works	<1%
23	Liang Xu, Yue Zhao, Xiaona Xu, Yigang Liu, Qiang Ji. "Latent Regressio... Crossref	<1%
24	University of Information Technology, Yangon on 2023-10-16 Submitted works	<1%
25	William Chan, Navdeep Jaitly, Quoc Le, Oriol Vinyals. "Listen, attend an... Crossref	<1%
26	ipfs.io Internet	<1%
27	proceedings.mlr.press Internet	<1%
28	mathworks.com Internet	<1%
29	Lecture Notes in Computer Science, 2016. Crossref	<1%
30	Patrick Doetsch, Mirko Hannemann, Ralf Schluter, Hermann Ney. "Inver... Crossref	<1%
31	Sogang University on 2020-08-13 Submitted works	<1%
32	University of Bradford on 2023-09-06 Submitted works	<1%

33	sersc.org Internet	<1%
34	unsworks.unsw.edu.au Internet	<1%
35	versologie.cz Internet	<1%
36	cse.iitd.ac.in Internet	<1%
37	isca-speech.org Internet	<1%