# Accented Speech Recognition

by

**Mohamed Mesto**

**Matriculation Number 390**

A thesis submitted to

Technische Universität Berlin
Faculty IV - Electrical Engineering and Computer Science
Institut für Softwaretechnik und Theoretische Informatik
Quality and Usability Lab

Master's / Bachelot's Thesis

January 16, 2023

Supervised by:
Prof. Dr. Sebastian Möller
Dr. Tim Polzehl

# Eidestattliche Erklärung / Statutory Declaration

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

I hereby declare that the thesis submitted is my own, unaided work, completed without any unpermitted external help. Only the sources and resources listed were used.

The independent and unaided completion of the thesis is affirmed by affidavit:

Berlin, January 16, 2023 Name

# Acknowledgments

First of all, I would like to thank ....

# Abstract

My Abstract ....

***Index Terms—*** Accented speech recognition, accent recognition, acoustic modeling, end-to-end ASR

# Zusammenfassung

summary in GERMAN

# Contents

# List of Figures

# List of Tables

# 1 Introduction

## 1.1 Motivation

In this paper, we will address the main engine in the field of neural networks[5, 6, 7] in general and automatic speech recognition in particular, which is transformers[8]. What contributed to her gaining this reputation is the multi-head self-attention approach, which instantly provides a high-handed position connection in parallel for the entire series, Instead of utilizing memory conditions to catch long-range dependencies in recurrent neural networks.

Nevertheless, the noteworthy transformer-based model architectures in automatic speech recognition (ASR) are Connectionist temporal classification (CTC)[9, 10], sequence-to- sequence [11, 12, 13, 14, 15], Neural transducer [16, 17, 18] and traditional hybrid systems[19, 20].

The results of the tests carried out on the public LibriSpeech showed a resounding success. The Emformer achieved a very low Word Error Rate (WER) in the clean-up test for a certain average latency [21]. Nevertheless, One of the promising technologies in this field is the Transformer Transducer. It exceeded the neural transducer with the Long-Short Term Memory (LSTM) or the Bidirectional Long-Short Term Memory (BLSTM) networks and achieved a surprising WER on the test-clean set and the test-other set.[17]. Moreover, One of the new technologies that grabbed the spotlight was the wav2vec2 pretrained model. It supported the ASR Systems in the Dysarthric speech recognition field[22].

The rest of this paper is organized as follows. Chapter 1, Chapter 2, Chapter 3, Chapter 4, Chapter 5

### 1.1.1 Improved/significantly improved

## 1.2 Background

text .... Wikipedia [1]

[1] https://www.wikipedia.org/

## 1.3 Problem Statement

## 1.4 Outline

# 2 Literature Review

## 2.1 Speech Recognition

During the last decades, the reliability of the speech recognition system has increased and its use in solving many industrial, medical and various challenges in different areas of daily life[23]. Automatic Speech Recognition (ASR), computer speech recognition, or Speech-To-Text (STT) are all terms for the concept of speech recognition. It is a sub-domain of the joint work between computer science, computer engineering, and computational linguistics. His work is summarized by approaches and procedures that allow the recognition of spoken language and its translation into text by computers[1]. In this paper, we will discuss some of these techniques, the most important of which are: Emformer Model [21, 5], DeepSpeech Model [24], Wav2vec2 [22], Transformer-Transducers [17], Conformer Model [18], Recurrent Neural Network Transducer (RNN-T)[16], and others. The following is a brief intro to each method as well as its advantages and disadvantages.

## 2.2 Acoustic Models

### 2.2.1 Non-streaming versus Streaming Models

Audio formats performed through the Web fall into two general classifications. The user must download non-streaming audio files to the user's hard disk before being able to play after that, unlike Streaming audio files, which can play almost immediately and continue playing while they are downloading. However, each category has advantages and disadvantages. In artificial intelligence algorithms specialized in dealing with linguistics, stream processing is a programming layout that considers data streams, or series of events in time, as the primary input and output entities of analysis. Parallel computing empowers the Stream processing systems for data streams and relies on streaming algorithms for efficient

[1] https://www.wikipedia.org/

performance[2].

An additional detailed discussion of specific streaming and non-streaming acoustic models follows.

### 2.2.2 AM-TRF Model

The Streaming Transformer with Augmented Memory (AM-TRF) [25] ...

**2.1a**

### 2.2.3 Emformer Model

Efficient Memory transformer based acoustic model for low latency streaming speech recognition is referred to simply as Emformer. It is one of the promising technologies in Accented Speech Recognition field. It earns this position because of the enhancement of the AM-TRF model. First, Emformer reduces the enumeration by extracting the repeated calculation from the left context block by caching the key and value in earlier segments' self-attention. Second, Emformer carries over the memory bank from the lower layer instead of passing the memory bank within the recent layer in AM-TRF motivated from the Transformer-XL features [5, 26]. Third, Emformer stops the resume vector's attention with the memory bank to evade overweighting the most left element of context information. Finally, the significant property of the Emformer for low latency speech recognition is presented in a parallelized block processing training strategy [21].

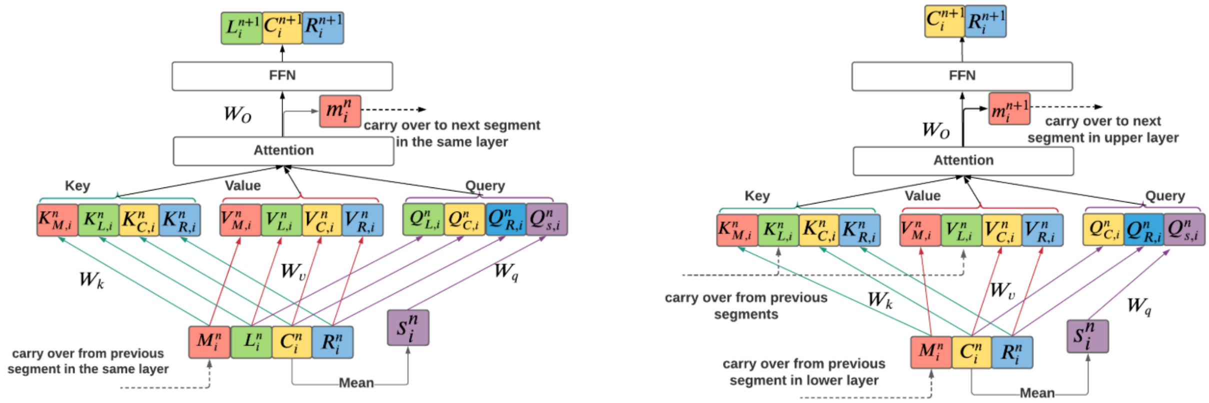
As shown in the previous paragraph, though, the AM-TRF has proven itself in the field of Accented Speech Recognition[25]. But his performance in the left context was unsatisfactory. Moreover, The main impetus behind the development of the Emformer technology is due to the shortcomings in implementing concatenated processing by AM-TRF. Figure **2.1b** shows one layer of the Emformer. The following subsections represent the significant advancements completed in Emformer.

- Cache key and value from previous segments

  In the AM-TRF model, Figure **2.1a** , the re-computation of $L_i^n$ for every step is required during the processing. Hence we need only to cache the projections from the earlier segments. The improvements of the Emformer are displayed in Figure **2.1b**. There is only demand to compute the key, and value projections for the memory bank, center, and right context. Besides, Since there is no need to give output from the left context block for the next layer, Emformer holds the computation of query projection of the left context.

  To understand the superiority of the Emformer over the AM-TRF, we will assume the following:

[2] https://www.wikipedia.org/




(a) AM-TRF

(b) Emformer

**Figure 2.1:** Comparison of AM-TRF with Emformer overlapping

[21]

M: is the length of the memory bank.

L: is the length of the left context block.

R: is the length of the right context.

C: is the length of the center context.

h: is the number of heads in the multi-head self-attention, and per head dimension is d.

Besides, the mean of the center segment That takes the value one is equal to the summary vector $S_i^n$. Nevertheless, the model dimension, dh, is substantially more enormous than any of M, L, R, and C. The implementation of the memory bank utilizes a ring buffer form with a short length. Hence, the Emfromer is more efficient than the AM-TRF by keeping around $\frac{L}{L+R+C}$ of AM-TRF computation [21].

- Carryover Memory Vector from Earlier Segments in the Lower Layer

  To benefit from the opportunities proposed by the Graphics processing unit (GPU), Emformer supports block processing in parallel during the training stage. It takes the memory bank input from previous segments in the lower layer instead of the same layer. Accordingly, the whole sequence is trained in parallel for each Emformer layer [21]. Unlikely to the AM-TRF Method. Whereas the auto-regression feature causes the block processing to be sequential.

- Prevent Attention between the Summary Vector with the Memory Bank

  To avoid the gradient disappearance or damage in the Emformer Technology. Moreover, to achieve a stable training accuracy for long-form speech, we need to allocate the attention weight between the summary vector and the memory bank to zero. At odds, embedding the memory bank information in the recent memory vector amplifies the most left context information [21].

- Look-Ahead Context Leaking handling

  During the training scenario, Emformer handles the input sequence quite parallel. Instead of physically dividing the input sequence into each layer, Emformer involves attention masks to limit the reception field [5, 27]. Unfortunately, this method has disadvantages represented in the look-ahead of context leaking. Adding to the actual size of the right context is when multiple transformer layers are stacked on top of each other. The right side of Figure **2.2** illustrates how the Emformer avoids the look-ahead context-leaking case during the training process. Emformer produces a hard copy of each segment's look-ahead context and sets the look-ahead context copy at the input sequence's start. To understand the concept, as we notice, the output in frame 5 in the first chunk only uses the information from the current piece jointly with frame 8 of the right context, excluding the right context leaking.

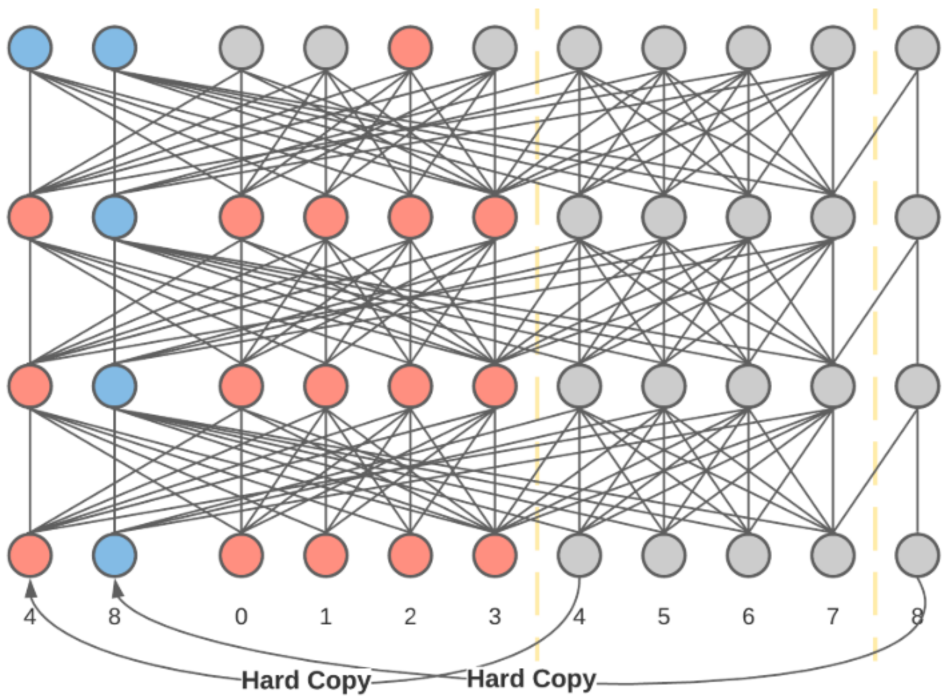



**Figure 2.2:** Representative of preventing look-ahead context leaking. Considering that: The chunk size value is 4, and the right context size is 1

[21]

### 2.2.4 Deepspeech Model

Alongside human life development, ASR systems' challenges are becoming very high due to the complicated architecture of deep neural networks (DNNs). Moreover, information systems that depend on audio inputs have integrated all the service devices surrounding us. Including IoT and personal identification systems in the home, Alexa[3] has a built-in ASR to comprehend our instructions. In our cars like summalinguae[4], workplace ,or PCs (e.g., to speak with Cortana[5] virtual assistant in Windows PC). Furthermore, in our mobile devices, there are ASRs that we can use to type texts or communicate with a mobile virtual assistant (e.g., Siri[6] and Google Assistant[7]). Which increased the risks of manipulating them on the lives of individuals or societies to an unimaginable extent [28, 29]. These manipulated inputs are called audio adversarial examples. The goal of these procedures is to deliver audio inputs different from the natural inputs to break into the property of others, such as accessing and controlling the work system of a company, house, or car. This misconception may risk the privacy of people [30, 31].

In this chapter, as a study use case, we will focus on, DeepSpeech, an End-to-End model for ASR. Furthermore, Deeply highlights a class of interpretability algorithms named attribution-based approaches [32, 33]. This strategy aims to estimate the DNN's most significant input's attributions to the output's behavior. Hence, we handle three visualization techniques: Shapley Additive Explanations (SHAP), Layer-wise Relevance Propagation (LRP), and Saliency Maps. We compare these methods and discuss possible applications, such as detecting adversarial examples [30, 34, 35].

The input of the DeepSpeech function is digital audio to generate the output "considerable possible" text transcript of that audio. Which grant it the name "automatically transcribing spoken audio tool." [8].

The mechanism of action of the DeepSpeech represented in taking a stream of audio as input and converting that stream of audio into a sequence of characters in the established alphabet [1]. Since there are two phases to be implemented. The audio is converted into a series of possibilities over characters in the alphabet in the first stage. Secondly, this series of options is converted into a string of characters. To achieve the first stage, we employ the Deep Neural Network. The DNN is trained on the audio and matchable text transcripts. Moreover, the neural model is qualified to predict the text from speech. Therefore, the first phase, the acoustic model, is well-deservedly called a phonetic transcriber. Besides, an N-gram language model is involved in accomplishing the second phase [36, 37]. The N-gram language model is trained on a text collection that is usually various from the text transcripts of the audio. Furthermore, The language model is trained to predict text from the previous text. In other words, we can consider the language model as a spelling and grammar checker.

**2.3**

[3] https://developer.amazon.com/en-GB/alexa
[4] https://summalinguae.com/language-technology/the-present-and-future-of-in-car-speech-recognition/
[5] https://www.microsoft.com/en-us/cortana/
[6] https://www.apple.com/siri/
[7] https://assistant.google.com/
[8] https://mozilla.github.io/deepspeech-playbook/

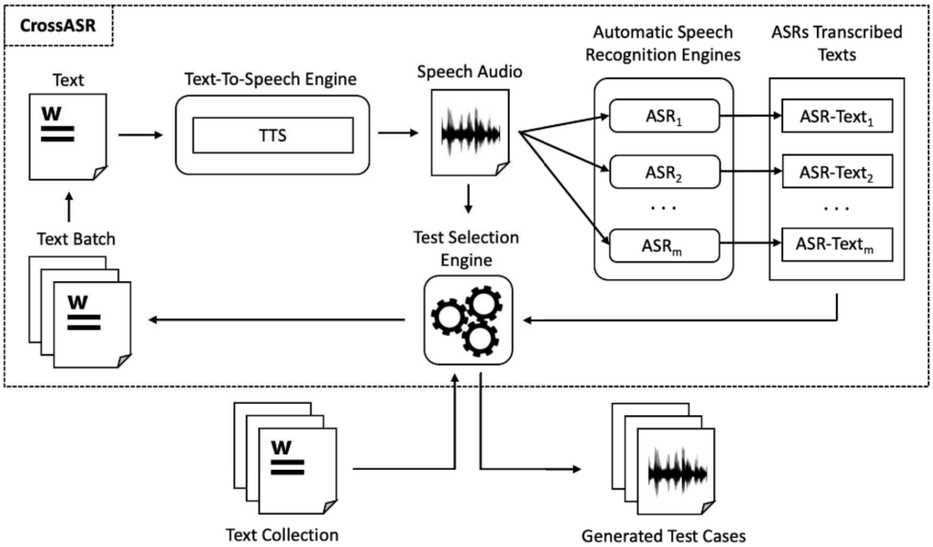
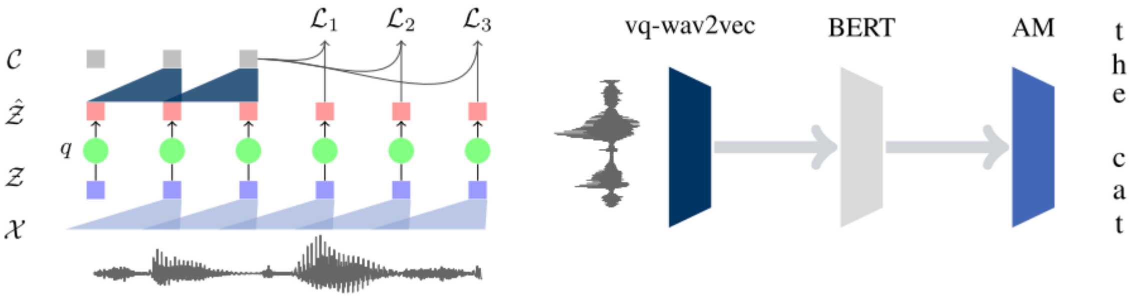



**Figure 2.3:** Architecture of CrossASR[1]

### 2.2.5 Wav2Vec Model

The feature's investigation of unsupervised pre-training for speech recognition will accomplish by the learning representations of unprocessed audio. Another promising Acoustic model [38] Wav2vec is trained on enormous amounts of unlabeled audio data. Moreover, the results are engaged to empower the acoustic model training. I.e., using a noise contrastive binary classification task to pre-train and optimize a basic multi-layer convolutional neural network. the outcomes of this method show positive improvements in the WER compared to Deepspeech 2 model. Wav2vec [38] learns representations of input audio data by addressing a self-supervised context-prediction assignment with the exact loss process as word2vec [39, 40]. In recent years, the spectrum of employment of learning discrete representations of speech has expanded in various fields [41, 42]. And most of the attention focused on automatic encoding to detect discrete units [43, 44, 45] taking two lanes; one way is to utilize predicting context information in order to learn continuous speech representations in a self- supervised mode [38, 46, 47]. The other way is united and correlated with an autoregressive model [48]. Nevertheless, recently, both ways of research combined. but rather than using rebuilding the input; learning discrete representations of speech via a context prediction task is employed [2]. This guided straight to involve a powerful performance of NLP algorithms to speech data which presented in Figure **2.4**.

From another angle, not only is the diversity of languages spoken by humanity, which exceeds seven thousand languages [49], the only obstacle, but the main challenge for speech recognition systems is to provide thousands of hours of written speech to reach sufficient performance. Regardless, in consider-




(a) vq-wav2vec

(b) Discretized speech training pipeline

(a) The vq-wav2vec encoder maps raw audio ($\mathcal{X}$) to a dense representation ($\mathcal{Z}$) which is quantized (q) to $\hat{\mathcal{Z}}$ and aggregated into context representations ($\mathcal{C}$); training requires future time step prediction. (b) Acoustic models are trained by quantizing the raw audio with vq-wav2vec, then applying BERT to the discretized sequence and feeding the resulting representations into the acoustic model to output transcriptions.

**Figure 2.4:** [2]

able scenes, labeled data is much harder to earn than unlabeled data. However, in speech recognition systems, Self-Supervised Learning relies on unlabeled samples to learn general data representations and to calibrate the model on labeled data. However, in speech recognition systems, Self-Supervised Learning relies on unlabeled samples to learn general data representations and to calibrate the model on labeled data. Hence, this approach has been victorious in the NLP domain [50, 51]. Likewise, in computer vision space [52, 53]. This guides us to the research effort [54] which shows a Framework for Self-Supervised Learning of Speech Representations. I.e., Wav2vec 2.0 [9]. Comparable to masked language modeling [6]; the method applies two stages: the first stage describes encoding speech audio through a multi-layer convolutional neural network. And the second stage describes masking spans of the resulting latent speech representations [55, 56].

The difficulties of dysarthric speech recognition to gain enough training data and a serious mismatch in speaker attributes was the motivation behind improving the recent Wav2vec2 to empower ASR technologies. Moreover, Feature space Maximum Likelihood Linear Regression (fMLLR) and x-vectors offer many advantages for dysarthric speech without hiring a huge amount of data. The scientific article [57] provides an approach, a flexible adaptation network, for fine-tuning Wav2vec2 employing fMLLR and x-vectors features.

[9] Code and models are available at https://github.com/pytorch/fairseq

### 2.2.6 Conformer Model

Recently, a Convolution-augmented Transformer for Speech Recognition is known as Conformer [18]. It passes successfully many exterminations of ASR. Nevertheless, The Conformer-Transducer (ConformerT) has proven its worth in several measurements [58, 59]. ConformerT becomes on top in E2E ASR systems; considering gains in its characteristics from conformer and transducer [59, 60]. Because of uniting the transformer and the convolution module in a parameter-efficient approach. Moreover, it has been and still is an active pursuit to obtain high accuracy when addressing Accented Speech Recognition through the development of the Conformer, which was offered for the first time in [3, 60]. This Acoustic model can be trained by utilizing E2E RNN-T loss [61] alongside a label encoder and a Conformer-based acoustic encoder (AEncoder). Towards making the word error rate as low as possible. Furthermore, to improve the Conformer-Transducer ASR system in a refined representation learning manner; the scientific paper [3] suggests a phonetic-assisted multi-target units (PMU) modeling approach. in more detail, the processing form passes in two stages: the first is summarized in employing PMU the pronunciation-assisted sub-word modeling (PASM). And the second stage poses employing byte pair encoding (BPE) to produce phonetic-induced and text-induced target units individually. At that moment, PMU, a paraCTC, and a pcaCTC are engaged to empower the acoustic encoder. additionally, those components merge the PASM and BPE units at different stations for Connectionist temporal classification (CTC) and transducer multi-task training. Figure **2.5** depicts the structure.

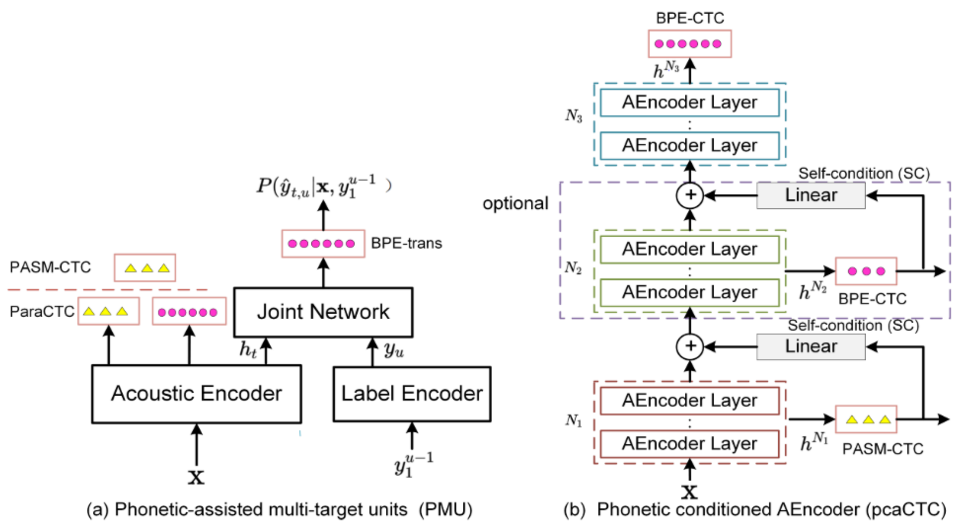
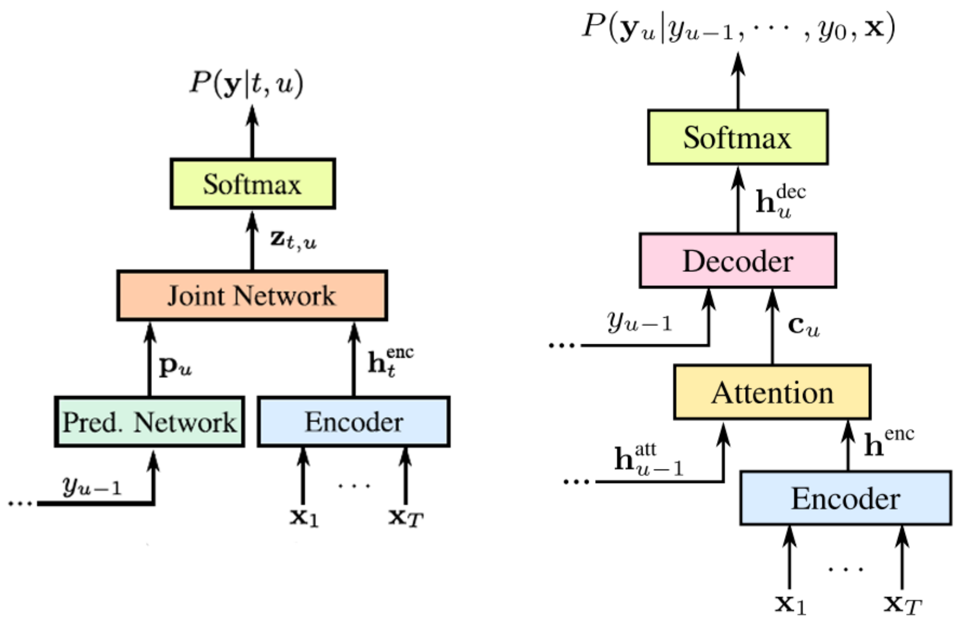



**Figure 2.5:** Structure of (a) the proposed phonetic-assisted multi-target units (PMU) and (b) its phonetic conditioned AEncoder (pcaCTC)[3]

### 2.2.7 RNN-T Model

The advance of ASR systems over time continues to be noteworthy [62]. Likewise, The quest to raise the efficiency of speech recognition systems in recent years [62] has made accented and linguistic models involved in the formation of hybrid speech recognition systems. However, it had obstacles that negatively affected the systems in terms of cost and the need for human intervention due to memory exhaustion and computational complicatedness, which makes it difficult to adopt these hybrid systems in cases with limited capabilities in terms of resources, as they are usually used in most personal computers or smartphones. Hence, there is an urgent need to develop end-to-end technologies [63, 64, 65]. at that instant, one of these mechanisms that grabbed the limelight because of their advantages such as adaptability, exactness, and performance is recurrent neural network transducer (RNN-T) [66]. in addition, Recurrent neural networks (RNNs) similarly long-short-term memory (LSTM) have met the requirements of sequence modeling in the ASR domain. Nevertheless, of the virtues worth noting, the recurrent connection from the previous state to the current state to propagate contextual information is powerful. but still has weak points. Such as the disability to compute in parallel. and capturing extended contexts is also challenging. This makes the door open to grant the attention mechanism [8, 67] the favorability for sequence modeling. This empowers the Transformer model to conclude state-of-the-art results in sequence-to-sequence responsibility [8].




(A) RNN-Transducer (B) Attention-based Model

**Figure 2.6:** Schematic diagrams of the RNN-T architecture (A) and the LAS architecture (B)[4].

As noted from the last example and many of the earlier models, Automatic speech recognition systems are often trained from typical speech. While the scientific paper [68] focuses on addressing the Accented speech recognition matter, and speech recognition issues from persons with amyotrophic lateral sclerosis (ALS). Normally the RNN-T utilizes two main elements: an encoder and decoder network, Which support him to act as a unidirectional model to conduct streaming ASR. However, the work [68, 4] use a bidirectional encoder without attention which shows an impressive performance. Figure **2.6** illustrates the RNN-T and LAS construction.

## 2.3 Language Models

### 2.3.1 N-gram Model

### 2.3.2 Transformer Model

### 2.3.3 GPTx Model

## 2.4 Streaming/non-streaming Models

# 3 Data Augmentation

## 3.1 Data Augmentation - Noise Factor

## 3.2 Data Augmentation - Time Factor

# 4 Experimentation

## 4.1 Dataset

## 4.2 Improved/ significantly improved

Data Augmentation (noise, time) –> librosa, specAug

************** That is what I would do as you could still introduce RTF, WER etc. in later sections and describe the results in better detail.

Will refine again

*********

The results of the tests carried out on the public LibriSpeech showed a resounding success. The Emformer achieved a WER of 2.50%in the clean-up test and 5.62% in the other-test for an average latency of 960ms [21]. Besides, at a low latency of 80 ms, Emformer gains WER 3.01% on test-clean and 7.09% on test-other. Nevertheless, One of the promising technologies in this field is the Transformer Transducer. It exceeded the neural transducer with LSTM/BLSTM networks and achieved a WER of 6.37% on the test-clean set and 15.30% on the test-other set[17]. Moreover, One of the new technologies that grabbed the spotlight was the wav2vec2 pretrained model. It supported the ASR Systems in the Dysarthric speech recognition field[22].

# 5 Case-Study / Design / Implementation

## 5.1 Using PYSA Data

using PYSA data and openly available data

# 6 Discussion

Evaluation using PYSA

# 7 Conclusion

text

## 7.1 Abbreviations

| Abbreviations | |
|---|---|
| **Abbreviations** | **Definition** |
| RNN-T | Recurrent Neural Network Transducer |
| TDNN | Time delay neural network |
| VGGNet | VGG stands for Visual Geometry Group; it is a standard deep Convolutional Neural Network (CNN) architecture with multiple layers |
| CNN | architecture with multiple |
| LSTM | Long-Short Term Memory |
| BLSTM | Bidirectional Long-Short Term Memory |
| WER | word error rates |
| fMLLR | Feature space Maximum Likelihood Linear Regression |
| RTF | real-time factor |
| AM-TRF | Augmented memory transformer |
| EMFORMER | Efficient memory transformer based acoustic model for low latency streaming speech recognition |
| FFN | Feed-Forward Network |
| LRP | Layer-wise Relevance Propagation |
| SHAP | Shapley Additive Explanations |
| WIT | Wave Inversion Technology |
| BERT | Deep Bidirectional Transformer |
| WSJ | Wall Street Journal |
| PMU | Phonetic-assisted multi-target units |
| PASM | Pronunciation-assisted sub-word modeling |
| BPE | byte pair encoding |
| pcaCTC | Phonetic Conditioned AEncoder CTC |
| CTC | Connectionist temporal classification |
| ALS | Amyotrophic Lateral Sclerosis |

**Table 7.1:** Accented Speech Recognition's Abbreviations

# Bibliography

[1] M. H. Asyrofi, F. Thung, D. Lo, and L. Jiang, "Crossasr: Efficient differential testing of automatic speech recognition via text-to-speech," in *2020 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 2020, pp. 640–650.

[2] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," *arXiv preprint arXiv:1910.05453*, 2019.

[3] L. Li, D. Xu, H. Wei, and Y. Long, "Phonetic-assisted multi-target units modeling for improving conformer-transducer asr system," *arXiv preprint arXiv:2211.01571*, 2022.

[4] R. Prabhavalkar, K. Rao, T. N. Sainath, B. Li, L. Johnson, and N. Jaitly, "A comparison of sequence-to-sequence models for speech recognition." in *Interspeech*, 2017, pp. 939–943.

[5] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[7] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer." *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[9] J. Salazar, K. Kirchhoff, and Z. Huang, "Self-attention networks for connectionist temporal classification in speech recognition," in *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2019, pp. 7115–7119.

[10] F. Zhang, Y. Wang, X. Zhang, C. Liu, Y. Saraf, and G. Zweig, "Faster, simpler and more accurate hybrid asr systems using wordpieces," *arXiv preprint arXiv:2005.09150*, 2020.

[11] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5884–5888.

[12] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang *et al.*, "A comparative study on transformer vs rnn in speech applications," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 449–456.

[13] M. Sperber, J. Niehues, G. Neubig, S. Stüker, and A. Waibel, "Self-attentional acoustic models," *arXiv preprint arXiv:1803.09519*, 2018.

[14] S. Zhou, L. Dong, S. Xu, and B. Xu, "Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin chinese," *arXiv preprint arXiv:1804.10752*, 2018.

[15] C. Wang, Y. Wu, S. Liu, J. Li, L. Lu, G. Ye, and M. Zhou, "Low latency end-to-end streaming speech recognition with a scout network," *arXiv preprint arXiv:2003.10369*, 2020.

[16] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7829–7833.

[17] C.-F. Yeh, J. Mahadeokar, K. Kalgaonkar, Y. Wang, D. Le, M. Jain, K. Schubert, C. Fuegen, and M. L. Seltzer, "Transformer-transducer: End-to-end speech recognition with self-attention," *arXiv preprint arXiv:1910.12977*, 2019.

[18] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[19] Y. Wang, A. Mohamed, D. Le, C. Liu, A. Xiao, J. Mahadeokar, H. Huang, A. Tjandra, X. Zhang, F. Zhang *et al.*, "Transformer-based acoustic modeling for hybrid speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6874–6878.

[20] D. Povey, H. Hadian, P. Ghahremani, K. Li, and S. Khudanpur, "A time-restricted self-attention layer for asr," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018,

pp. 5874–5878.

[21] Y. Shi, Y. Wang, C. Wu, C.-F. Yeh, J. Chan, F. Zhang, D. Le, and M. Seltzer, "Emformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6783–6787.

[22] M. Karthick Baskar, T. Herzig, D. Nguyen, M. Diez, T. Polzehl, L. Burget, and J. Honza"Černockỳ, "Speaker adaptation for wav2vec2 based dysarthric asr," *arXiv e-prints*, pp. arXiv–2204, 2022.

[23] J. Benesty, J. Chen, and Y. Huang, "Automatic speech recognition: A deep learning approach," 2008.

[24] K. Markert, R. Parracone, M. Kulakov, P. Sperl, C.-Y. Kao, and K. Böttinger, "Visualizing Automatic Speech Recognition – Means for a Better Understanding?" in *Proc. 2021 ISCA Symposium on Security and Privacy in Speech Communication*, 2021, pp. 14–20.

[25] C. Wu, Y. Wang, Y. Shi, C.-F. Yeh, and F. Zhang, "Streaming transformer-based acoustic models using self-attention with augmented memory," *arXiv preprint arXiv:2005.08042*, 2020.

[26] L. Lu, C. Liu, J. Li, and Y. Gong, "Exploring transformers for large-scale speech recognition," *arXiv preprint arXiv:2005.09684*, 2020.

[27] X. Chen, Y. Wu, Z. Wang, S. Liu, and J. Li, "Developing real-time streaming transformer transducer for speech recognition on large-scale dataset," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5904–5908.

[28] S. Hu, X. Shang, Z. Qin, M. Li, Q. Wang, and C. Wang, "Adversarial examples for automatic speech recognition: Attacks and countermeasures," *IEEE Communications Magazine*, vol. 57, no. 10, pp. 120–126, 2019.

[29] H. Abdullah, K. Warren, V. Bindschaedler, N. Papernot, and P. Traynor, "Sok: The faults in our asrs: An overview of attacks against automatic speech recognition and speaker identification systems," in *2021 IEEE symposium on security and privacy (SP)*. IEEE, 2021, pp. 730–747.

[30] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *2018 IEEE security and privacy workshops (SPW)*. IEEE, 2018, pp. 1–7.

[31] R. Mitev, A. Pazii, M. Miettinen, W. Enck, and A.-R. Sadeghi, "Leakypick: Iot audio spy detector," in *Annual Computer Security Applications Conference*, 2020, pp. 694–705.

[32] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," *arXiv preprint arXiv:1711.06104*, 2017.

[33] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[34] Y. Qin, N. Carlini, G. Cottrell, I. Goodfellow, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *International conference on machine learning*. PMLR, 2019, pp. 5231–5240.

[35] T. Dörr, K. Markert, N. M. Müller, and K. Böttinger, "Towards resistant audio adversarial examples," in *Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence*, 2020, pp. 3–10.

[36] K. Markert, D. Mirdita, and K. Böttinger, "Language dependencies in adversarial attacks on speech recognition systems," *arXiv preprint arXiv:2202.00399*, 2022.

[37] P. Michel, X. Li, G. Neubig, and J. M. Pino, "On evaluation of adversarial perturbations for sequence-to-sequence models," *arXiv preprint arXiv:1903.06620*, 2019.

[38] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[39] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[40] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[41] M. Versteegh, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015: Proposed approaches and results," *Procedia Computer Science*, vol. 81, pp. 67–72, 2016.

[42] E. Dunbar, R. Algayres, J. Karadayi, M. Bernard, J. Benjumea, X.-N. Cao, L. Miskic, C. Dugrain, L. Ondel, A. W. Black *et al.*, "The zero resource speech challenge 2019: Tts without t," *arXiv preprint arXiv:1904.11469*, 2019.

[43] A. Tjandra, B. Sisman, M. Zhang, S. Sakti, H. Li, and S. Nakamura, "Vqvae unsupervised unit discovery and multi-scale code2spec inverter for zerospeech challenge 2019," *arXiv preprint arXiv:1905.11449*, 2019.

[44] R. Eloff, A. Nortje, B. van Niekerk, A. Govender, L. Nortje, A. Pretorius, E. Van Biljon, E. van der Westhuizen, L. van Staden, and H. Kamper, "Unsupervised acoustic unit discovery for speech synthesis using

discrete latent-variable neural networks," *arXiv preprint arXiv:1904.07556*, 2019.

[45] J. Chorowski, R. J. Weiss, S. Bengio, and A. Van Den Oord, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 12, pp. 2041–2053, 2019.

[46] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[47] Y.-A. Chung and J. Glass, "Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech," *arXiv preprint arXiv:1803.08976*, 2018.

[48] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," *arXiv preprint arXiv:1904.03240*, 2019.

[49] M. P. Lewis, G. F. Simons, and C. D. Fennig, "Ethnologue: languages of the world, dallas, texas: Sil international," *Online version: http://www. ethnologue. com*, vol. 12, no. 12, p. 2010, 2009.

[50] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, and K. Lee, "Luke 440 zettlemoyer. deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, vol. 441, 2018.

[51] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[52] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[53] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[54] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[55] D. Jiang, X. Lei, W. Li, N. Luo, Y. Hu, W. Zou, and X. Li, "Improving transformer-based speech recognition using unsupervised pre-training," *arXiv preprint arXiv:1910.09932*, 2019.

[56] W. Wang, Q. Tang, and K. Livescu, "Unsupervised pre-training of bidirectional speech encoders via masked reconstruction," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal*

*Processing (ICASSP)*. IEEE, 2020, pp. 6889–6893.

[57] M. K. Baskar, T. Herzig, D. Nguyen, M. Diez, T. Polzehl, L. Burget, and J. H. Černocký, "Speaker adaptation for wav2vec2 based dysarthric asr," 2022. [Online]. Available: https://arxiv.org/abs/2204.00770

[58] F. Boyer, Y. Shinohara, T. Ishii, H. Inaguma, and S. Watanabe, "A study of transducer based end-to-end asr with espnet: Architecture, auxiliary loss and decoding strategies," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 16–23.

[59] J. Li, Y. Wu, Y. Gaur, C. Wang, R. Zhao, and S. Liu, "On the comparison of popular end-to-end models for large scale speech recognition," *arXiv preprint arXiv:2005.14327*, 2020.

[60] W. Huang, W. Hu, Y. T. Yeung, and X. Chen, "Conv-transformer transducer: Low latency, low frame rate, streamable end-to-end speech recognition," *arXiv preprint arXiv:2008.05750*, 2020.

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

[62] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 30–42, 2011.

[63] A. Mohamed, D. Okhonko, and L. Zettlemoyer, "Transformers with convolutional context for asr," *arXiv preprint arXiv:1904.11660*, 2019.

[64] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*. PMLR, 2014, pp. 1764–1772.

[65] L. Dong, F. Wang, and B. Xu, "Self-attention aligner: A latency-control end-to-end model for asr using self-attention network and chunk-hopping," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5656–5660.

[66] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang *et al.*, "Streaming end-to-end speech recognition for mobile devices," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6381–6385.

[67] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *Advances in neural information processing systems*, vol. 28, 2015.

[68] J. Shor, D. Emanuel, O. Lang, O. Tuval, M. Brenner, J. Cattiau, F. Vieira, M. McNally, T. Charbonneau,

M. Nollstadt *et al.*, "Personalizing asr for dysarthric and accented speech with limited data," *arXiv preprint arXiv:1907.13511*, 2019.

[69] T. Polzehl, B. Naderi, F. Köster, and S. Möller, "Robustness in speech quality assessment and temporal training expiry in mobile crowdsourcing environments," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[70] A. C. Rouse, "A preliminary taxonomy of crowdsourcing," *ACIS 2010 Proceedings*, vol. 76, pp. 1–10, 2010.

[71] K. Markert, D. Mirdita, and K. Böttinger, "Language Dependencies in Adversarial Attacks on Speech Recognition Systems," in *Proc. 2021 ISCA Symposium on Security and Privacy in Speech Communication*, 2021, pp. 25–31.

[72] J. Castro, D. Gómez, and J. Tejada, "Polynomial calculation of the shapley value based on sampling," *Computers & Operations Research*, vol. 36, no. 5, pp. 1726–1730, 2009.

[73] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[74] C.-T. Do and Y. Stylianou, "Weighting Time-Frequency Representation of Speech Using Auditory Saliency for Automatic Speech Recognition," in *Proc. Interspeech 2018*, 2018, pp. 1591–1595.

[75] L. Schulth, C. Berghoff, and M. Neu, "Detecting backdoor poisoning attacks on deep neural networks by heatmap clustering," *arXiv preprint arXiv:2204.12848*, 2022.

[76] W. Ge, M. Todisco, and N. Evans, "Explainable Deepfake and Spoofing Detection: An Attack Analysis Using SHapley Additive exPlanations," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 70–76.

[77] J. Pan, J. Shapiro, J. Wohlwend, K. J. Han, T. Lei, and T. Ma, "ASAPP-ASR: Multistream CNN and Self-Attentive SRU for SOTA Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 16–20.

[78] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[79] A. Agarwal and T. Zesch, "German end-to-end speech recognition based on deepspeech." in *KONVENS*, 2019.

[80] B. Alibegović, N. Prljača, M. Kimmel, and M. Schultalbers, "Speech recognition system for a service robot-a performance evaluation," in *2020 16th International Conference on Control, Automation, Robotics and Vision*

*(ICARCV)*. IEEE, 2020, pp. 1171–1176.

[81] M. Kleinert, N. Venkatarathinam, H. Helmke, O. Ohneiser, M. Strake, and T. Fingscheidt, “Easy adaptation of speech recognition to different air traffic control environments using the deepspeech engine,” 2021.

[82] S. G. Manepalli, D. Whitenack, and J. Nemecek, “Dyn-asr: Compact, multilingual speech recognition via spoken language and accent identification,” in *2021 IEEE 7th World Forum on Internet of Things (WF-IoT)*. IEEE, 2021, pp. 830–835.

[83] M. Zeineldeen, J. Xu, C. Lüscher, R. Schlüter, and H. Ney, “Improving the training recipe for a robust conformer-based hybrid model,” *arXiv preprint arXiv:2206.12955*, 2022.

[84] K. Deng, S. Cao, and L. Ma, “Improving accent identification and accented speech recognition under a framework of self-supervised learning,” *arXiv preprint arXiv:2109.07349*, 2021.

[85] X. Shi, F. Yu, Y. Lu, Y. Liang, Q. Feng, D. Wang, Y. Qian, and L. Xie, “The accented english speech recognition challenge 2020: Open datasets, tracks, baselines, results and methods,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6918–6922.

[86] M. Zeineldeen, J. Xu, C. Lüscher, W. Michel, A. Gerstenberger, R. Schlüter, and H. Ney, “Conformer-based hybrid asr system for switchboard dataset,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7437–7441.

[87] K. Rao, H. Sak, and R. Prabhavalkar, “Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer,” in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 193–199.

[88] T. Viglino, P. Motlicek, and M. Cernak, “End-to-end accented speech recognition.” in *Interspeech*, 2019, pp. 2140–2144.

[89] K. Rao, H. Sak, and R. Prabhavalkar, “Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer,” in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 193–199.

[90] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang, Q. Liang, D. Bhatia, Y. Shangguan, B. Li, G. Pundak, K. C. Sim, T. Bagby, S.-y. Chang, K. Rao, and A. Gruenstein, “Streaming end-to-end speech recognition for mobile devices,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6381–6385.

[91] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*. PMLR, 2016, pp. 173–182.

[92] A. Hinsvark, N. Delworth, M. Del Rio, Q. McNamara, J. Dong, R. Westerman, M. Huang, J. Palakapilly, J. Drexler, I. Pirkin *et al.*, "Accented speech recognition: A survey," *arXiv preprint arXiv:2104.10747*, 2021.

[93] N. Das, S. Bodapati, M. Sunkara, S. Srinivasan, and D. H. Chau, "Best of both worlds: Robust accented speech recognition with adversarial transfer learning," *arXiv preprint arXiv:2103.05834*, 2021.

[94] X. Shi, F. Yu, Y. Lu, Y. Liang, Q. Feng, D. Wang, Y. Qian, and L. Xie, "The accented english speech recognition challenge 2020: open datasets, tracks, baselines, results and methods," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6918–6922.

[95] M. Zeineldeen, J. Xu, C. Lüscher, W. Michel, A. Gerstenberger, R. Schlüter, and H. Ney, "Conformer-based hybrid asr system for switchboard dataset," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7437–7441.