DOT/FAA/TC-23/47

Federal Aviation Administration
William J. Hughes Technical Center
Aviation Research Division
Atlantic City International Airport
New Jersey 08405

Review of Automatic Speech Recognition Methodologies

August, 2023

Final report

U.S. Department of Transportation
Federal Aviation Administration

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof. The U.S. Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this report. The findings and conclusions in this report are those of the author(s) and do not necessarily represent the views of the funding agency. This document does not constitute FAA policy. Consult the FAA sponsoring organization listed on the Technical Documentation page as to its use.

This report is available at the Federal Aviation Administration William J. Hughes Technical Center's Full-Text Technical Reports page: actlibrary.tc.faa.gov in Adobe Acrobat portable document format (PDF).

Technical Report Documentation Page

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

1. Report No. DOT/FAA/TC-23/47	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle Review of Automatic Speech Recognition Methodologies		5. Report Date August 2023	
		6. Performing Organization Code	
7. Author(s) Gabriel Achour (GT), Oojas Salunke (GT), Alexia Payan (GT), Evan Harrison (GT) Chiheb Sahbani (Rowan), Giuseppina Carannante (Rowan), Gregory Ditzler (Rowan), Nidhal Bouaynaya (Rowan)		8. Performing Organization Report No.	
9. Performing Organization Name and Address Aerospace Systems Design Laboratory The Daniel Guggenheim School of Aerospace Engineering Georgia Institute of Technology 620 Cherry Street, Atlanta, GA 30332 Rowan University 201 Mullica Hill Rd Glassboro, NJ 08028		10. Work Unit No. (TRAIS)	
		11. Contract or Grant No. DTFACT-14-D-00004 692M152240001	
12. Sponsoring Agency Name and Address FAA Mike Monroney Aeronautical Center Flight Technologies and Procedures Division Flight Research & Analysis Branch, AFS-430 6500 S MacArthur Blvd, Bldg 26, Oklahoma City, OK 73169-6918		13. Type of Report and Period Covered	
		14. Sponsoring Agency Code AFS-430	
15. Supplementary Notes The FAA William J. Hughes Technical Center Aviation Research Division Technical Monitor was Huasheng Li			
16. Abstract This report highlights the crucial role of Automatic Speech Recognition (ASR) techniques in enhancing safety for air traffic control (ATC) in terminal environments. ASR techniques facilitate efficient and accurate transcription of verbal communications, reducing the likelihood of errors. The report also details the evolution of ASR technologies, converging to machine learning approaches from Hidden Markov Models (HMMs), Deep Neural Networks (DNNs) to End-to-End models. Finally, the report details the latest advancements in ASR techniques, focusing on transformer-based models that have outperformed traditional ASR approaches and achieved state-of-the-art results on ASR benchmarks.			
17. Key Words Machine Learning - Automatic Speech Recognition Machine Learning - Speech to Text Machine Learning - Natural Language Processing	18. Distribution Statement This document is available to the U.S. public through the National Technical Information Service (NTIS), Springfield, Virginia 22161. This document is also available from the Federal Aviation Administration William J. Hughes Technical Center at actlibrary.tc.faa.gov.		
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 47	22. Price

Contents

Figures

Tables

Acronyms

Acronym	Definition
ASDEX	Advanced Surveillance Data Exchange
ADS-B	Automatic Dependent Surveillance-Broadcast
ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
ATC	Air Traffic Control
BRNN	Bidirectional Recurrent Neural network
CAASD	Corporation's Center for Advanced Aviation System Development
CBP	Contextual Block Processing
CNN	Convolutional Neural Networks
CROPD	Closed Runway Operation Prevention Device
CTC	Connectionist Temporal Classification
DNN	Deep Neural Network
ETMS	En Route Traffic Management
FAA	Federal Aviation Administration
FAS	Flight Analysis System
GMM	Gaussian Mixture Model
GPU	Graphics Processing Unit
HMM	Hidden Markov Model
LSTM	Long Short Term Memory
NAS	National Airspace System
NOTAMs	Notices to Airmen
NOP	National Operations Portal
PBWP	Product Based Work Plan
PIREPs	Pilot Reports
NLP	Natural Language Processing
ReLu	Rectified Linear
RNN	Recurrent Neural Networks
RNN-T	Recurrent Neural Networks with Transducers
SVMs	Support Vector Machines
WSJ	Wall Street Journal

Executive summary

Automatic speech recognition (ASR) techniques are crucial in improving safety for Air Traffic Control (ATC) in terminal environments. In these environments, ATC controllers need to communicate with multiple pilots simultaneously, and any miscommunication or misunderstanding could have severe consequences. ASR techniques allow for more efficient and accurate transcription of verbal communications, reducing the likelihood of errors and misunderstandings. This can ultimately improve safety by ensuring that all parties have a clear understanding of the information being exchanged and can respond appropriately.

This report starts by discussing the MITRE Corporation's previous research on ASR, which focused on improving voice recognition for ATC systems. It then moves on to detail the evolution of ASR technologies to more modern machine learning approaches.

The development of ASR technologies covers key milestones, such as the introduction of Hidden Markov Models (HMMs) and Deep Neural Networks (DNNs), which have led to significant improvements in ASR accuracy, and ultimately lead to end-to-end techniques.

Finally, the report dives into the latest advancements in ASR techniques, specifically the use of transformer-based models. These models have achieved state-of-the-art results on a range of ASR benchmarks and have been shown to outperform traditional ASR approaches.

1 Introduction

Improving speech recognition in pilot Air Traffic Control (ATC) communication can have a significant impact on safety in aviation. Accurate and timely communication between pilots and ATC is critical for safe and efficient air traffic management. Miscommunication or misunderstandings between pilots and ATC can result in incidents, accidents, and even fatalities. Improving the accuracy and efficiency of speech recognition in ATC can ultimately enhance safety in aviation (Geacăr, 2010). The application of automatic speech recognition (ASR) in ATC has been explored by various research teams and demonstrated using different techniques (Helmke, Ohneiser, Muhlhausen, & Wies, 2016; Lin, et al., 2019; Lin, Guo, Zhang, Chen, & Yang, 2021)

This report explores the model components and their evolution that make up ASR. ASR techniques have been studied since the 1970s (Reddy, 1976) and are still today the interest of many research teams (Nassif, Shahin, Attili, Azzeh, & Shaalan, 2019). Specifically, this report examines the evolution of ASR methods, including Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM), and the modifications made to the HMM pipeline that ultimately led to the emergence of end-to-end models.

The report also showcases the potential of ASR technology in enhancing safety and efficiency within the National Airspace System (NAS). Focusing on the collaborative efforts of the MITRE Corporation's Center for Advanced Aviation System Development (CAASD) and the Federal Aviation Administration (FAA) in implementing advanced speech recognition technologies, such as the Flight Analysis System (FAS) and the Closed Runway Operation Prevention Device (CROPD), are highlighted to analyze the NAS and accurately identify potential hazards. The critical role that ASR technology plays in ensuring the highest level of safety for all airspace users is also discussed.

Moreover, the report highlights the advancements in deep learning (DL) techniques that have resulted in significant improvements in ASR accuracy and performance, including the emergence of Transformer models as a powerful alternative to RNN for ASR tasks. The report explores how these models leverage self-attention mechanisms to capture long-range dependencies and complex patterns in speech data, enabling more accurate and efficient recognition of speech. Furthermore, the report examines how Transformer models can be fine-tuned and adapted to various tasks, languages, and domains with relative ease, further enhancing the capabilities and performance of ASR models.

2 Past FAA and MITRE reports

The MITRE Corporation's CAASD has been working on the development of a state-of-the-art FAS as part of their Product Based Work Plan (PBWP) (McGuire & Feerrar, 2014). The FAS is designed to seamlessly merge and amalgamate large datasets from various sources such as the National Operations Portal (NOP), Airport Surface Detection Equipment Model X (ASDE-X), En Route Traffic Management System (ETMS), and Automatic Dependent Surveillance-Broadcast (ADS-B) (McGuire & Feerrar, 2014) . The system incorporates aircraft trajectory information with weather patterns, airspace procedures, and Notices to Airmen (NOTAMs) to create a comprehensive and exhaustive analysis (McGuire & Feerrar, 2014) . In addition to these features, the FAS also harnesses the power of cutting-edge ASR technology to integrate pilot-controller voice communications and undertake large-scale analyses. In this work, an extensive voice processing pipeline is outlined, with ASR being the first component. A significant part of the report details the context incorporation and semantic extraction which is required to merge the voice data with other data fields (McGuire & Feerrar, 2014) Later applications of the methodologies will show that ASR is not the only point of failure that impacts the overall system performance.

A major research focus of the MITRE CAASD team was the CROPD, which led to the identification of an extensive suite of applications aimed at early detection of surface safety events (Chen, et al., 2016). The research aimed to enhance and amplify the already robust speech and language processing capabilities, with two main technical objectives: first, identifying other potential applications for detecting surface safety events, and second, bolstering speech recognition performance on Tower controller audio (Chen, et al., 2016).

The use of ASR technology has demonstrated unparalleled potential for analyzing the NAS (Kopald H. , 2017). The research showed a 99% classification agreement between the algorithm and human-transcribed text in identifying missed approach initiators, thereby establishing the feasibility and efficacy of utilizing ASR for this critical task (Kopald H. , 2017). Furthermore, key phrases were recognized with remarkable 96% accuracy, and using simple logic, the initiator could be identified an impressive 90% of the time (Kopald H. , 2017).

The FAA gave the responsibility of researching speech recognition in the NAS to MITRE CAASD, with a focus on runway safety and investigating the potential benefits of speech recognition technology across the NAS (Kopald & Chen, 2019). The researchers developed advanced concepts for detecting erroneous surface operations in real-time, with low false positive rates but concerning false negative rates (Kopald & Chen, 2019).

In accordance with the MITRE CAASD's investigation into automated speech recognition for the FAA to support the NAS (Kopald, Chong, & Shepley, 2018), research has been conducted into the recognition of voice and the improvement of its performance in detecting safe clearances for closed runways. The study was aimed at providing a deeper understanding of the technology and its potential applications, with the FAA seeking a more comprehensive range of applications for speech recognition technology. This research is a part of a larger FAA portfolio for NAS safety, which has contributed to a more nuanced comprehension of the technology and its use cases, informing future FAA decisions (Kopald, Chong, & Shepley, 2018).

The MITRE/CAASD designed the Late or Missing Landing Clearance Detection and Notification System prototype as part of an FAA initiative to enhance aviation safety (Tarakan, 2012). The prototype utilizes automatic speech recognition, data fusion, and other technologies, and was hosted and demonstrated in the CAASD IDEA Lab (Tarakan, 2012). The system performed excellently in a simulated environment representing Hartsfield-Jackson Atlanta International Airport and is now ready for field customization and site-specific enhancements (Tarakan, 2012). ASR performance depends on audio fidelity, model conformity to realistic field characteristics, and additional airport/tower-specific adaptations necessary for optimization (Tarakan, 2012).

In the context of airspace security operations, ideal transcription conditions include low noise, single speaker, known speaker, and structured vocabulary. High-quality input is crucial for obtaining accurate transcripts, and investigation into microphone options such as noise-canceling microphones is recommended (Henriques, 2009). Storing transcribed text and lower quality recordings can save space without sacrificing audible sound quality (Henriques, 2009). The currently used codec, GSM 06.10, leads to a high Word-Error Rate; it is recommended to use the G.711 codec for better speech recognition accuracy (Henriques, 2009). To further increase accuracy, it is advised to record the DEN HQ separately, separate participants into individual tracks, and apply speech-transcription technology individually to each track (Henriques, 2009).

The MITRE CAASD's investigation highlights the potential and significance of speech recognition technology as part as a processing pipeline including semantic parsing in improving safety and efficacy within the National Airspace System. By assimilating ASR into various applications, such as FAS and CROPD, and by continuously refining the technology, the FAA can more effectively manage and monitor aviation operations while guaranteeing a high level of safety for all airspace users. Additional MITRE reports that were investigated in regard to speech-to-text are summarized with pros and cons in Appendix A.

3 Automatic Speech Recognition methods

In a traditional ASR system, language models and acoustic models are two key components that work together to transcribe spoken words into text. The typical pipeline is shown in Figure 1 (Renkens, 2017).

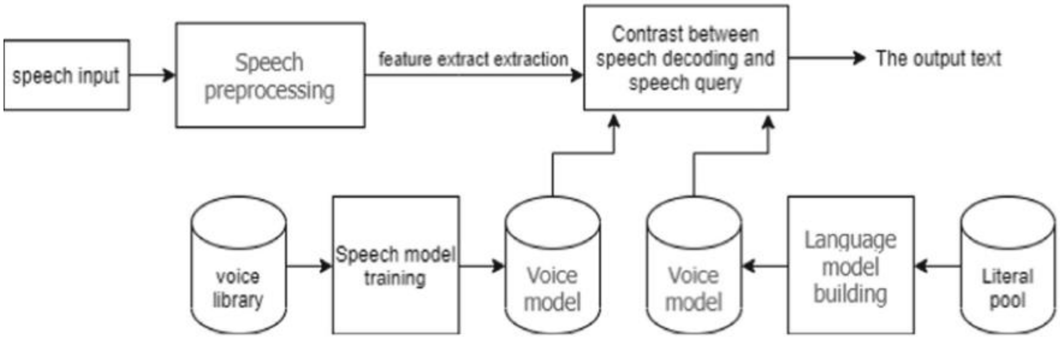

Figure 1. Traditional automatic speech recognition pipeline .

Acoustic models are responsible for mapping acoustic features extracted from the speech signal to probabilities of speech sounds, which are often represented as phonemes or sub-phonetic units. Acoustic models are typically trained using a large dataset of speech recordings and their corresponding transcriptions, using techniques such as Hidden Markov Models (HMMs) or Deep Neural Networks (DNNs).

Language models, on the other hand, are responsible for assigning probabilities to sequences of words. These models are trained on large text dataset and estimate the likelihood of observing a particular sequence of words based on the statistical patterns they have learned from the training data.

In an ASR system, the output of the acoustic model and the language model are combined. Specifically, the language model assigns probabilities to possible word sequences based on the statistical patterns of the training data, and the acoustic model assigns probabilities to possible speech sound sequences based on the characteristics of the acoustic features extracted from the speech signal. These probabilities are then combined to generate a transcription that is most likely given the observed speech signal.

The language model and acoustic model work together to improve the accuracy of the transcription. By considering the context of the spoken words, the language model can help to distinguish words that sound similar, while the acoustic model can help differentiate words that are phonetically similar.

In the literature, there exist many different ASR classifiers such as Support Vector Machines (SVM), Artificial Neural Networks (ANN) or k-NN classifiers. Yet, traditionally, ASR models have been leveraging Hidden Markov Gaussian Mixture Models (HMM-GMM) relying on probabilistic methods to compute the likelihood of a text string corresponding to an input sound wave (El Ayadi, Kamel, & Karray, 2011). This ASR technique has been around for decades and evolved as computational capabilities improved over the years. Progressively, DL models replaced the pipeline elements of HMM-GMM models and converged into end-to-end models.

3.1 Hidden Markov model and Gaussian Mixture models

The HMM aims to extract speech data from a noisy environment. It uses densities of probability assembled by a mixture of Gaussian functions. In the initial stage of the training process, a model is computed for each user and saved in the database. Later, when the user's spectral features are generated, the system searches the database to identify the model that most accurately matches these features. (Rodríguez, García-Crespo, & García, 1997).

Once the training data has been collected, the HMM-GMM model is built. The model consists of a set of HMMs, each corresponding to a different speech sound, and a set of GMMs, which model the probability distribution of the acoustic features associated with each speech sound. To transcribe new speech, the HMM-GMM model is used to decode the acoustic features extracted from the speech signal. The model calculates the probability of each speech sound given the acoustic features, and the most likely sequence of speech sounds is selected as the transcription. (Povey, 2004).

To perform automatic speech recognition, the conventional approach requires establishing a connection between the speech signal and the digital model during the front-end speech preprocessing stage. The combination of sampled speech signals is then used to predict the signal using the linear prediction analysis method. However, due to the complexities involved in extracting speech information, adapting the preprocessing models in traditional speech recognition to different scenarios characterized by diverse pronunciations of individuals speaking different languages, gender, and age becomes a challenging task (Renkens, 2017).

3.2 Modifications of the HMM pipeline

The HMM pipeline can be improved depending on the application, such as boosting by lattice composition to enhance the prediction accuracy of the contextual information such as a call sign for a ATC and pilot communication (Kocour, et al., 2021).

DL methods have significantly improved the accuracy of ASR systems by learning and modeling complex relationships between acoustic features and spoken language, allowing speech recognition systems to increase recognition rates substantially. One key advantage of DL methods in ASR is their ability to learn more robust representations of speech data, which allows them to capture complex patterns and variability in speech signals. By using large-scale datasets and complex neural architectures, DL models can learn to recognize more subtle and abstract speech features, such as tone, emphasis, and emotion.

During the initial stages of integrating DL with automatic speech recognition, researchers incorporated DL models into the framework of HMM/GMM. By using DL models to optimize the data for automatic speech recognition, they aimed to improve the performance of the system.

Adding neural networks to HMM/GMM models has been shown to significantly improve the accuracy of ASR systems. One approach that has proven particularly effective is the tandem approach, which combines a neural network-based acoustic model with a traditional HMM-based language model.

In the Tandem approach, the neural network-based acoustic model is used to extract high-level features from the speech signal, which are then used to train a traditional HMM-based language model. The resulting system can then be used to recognize speech with higher accuracy than traditional HMM/GMM models alone. (Vinalys & Ravuri, 2011).

Researchers have experimented with various approaches to incorporate DL models into existing components of ASR systems, with the goal of improving accuracy and robustness. One such approach is the DNN-HMM hybrid approach (Dahl, Member, Deng, & Acero, 2011), in which DL models are used to replace the GMM structure in the HMM. This approach allows for the use of more powerful neural network-based models to capture complex patterns and relationships in the speech data. These approaches have shown promising results and have paved the way for the development of more sophisticated and effective ASR systems that can better handle the challenges of real-world speech recognition applications.

Later, researchers tried to replace core elements of the traditional pipeline such as the language model or the acoustic model with DL models, such as the replacement of the acoustic model with a RNN architecture (Maas, et al., 2012). HMMs have been the gold standard in ASR for years, specifically when applied in an ATC environment (Ferreiros, et al., 2012). While HMMs have been successful in ATC, recent advances in DL, specifically end-to-end models such as Transformers, have shown promise in improving speech recognition accuracy, and are being explored for use in ATC. The trend over the years was to progressively introduce DL models in the prediction pipeline, to ultimately converge to one DL architecture encapsulating the entire pipeline, End-to-End models.

3.3 End-to-end models

One issue with hybrid systems is that several intermediate models (acoustic model, language model, lexicon) either need expert linguistic knowledge or be trained and designed separately. In the last few years, there has been an increasing focus on the development and adoption of end-to-end systems in ASR.

End-to-end models are considered better for ASR because they offer a simpler, more efficient, and more accurate approach to speech recognition compared to traditional ASR systems. End-to-end models are neural network-based models that can take raw speech signal as input and output the corresponding transcript directly, without the need for separate components for feature extraction, acoustic modeling, and language modeling.

By eliminating the need for separate components, end-to-end models reduce the complexity of ASR systems, resulting in faster training, reduced inference times, and lower computational requirements. Additionally, end-to-end models can learn more effectively from data, leading to better accuracy and robustness, especially in scenarios where the speech data is highly variable and noisy.

Furthermore, end-to-end models are more flexible and adaptable than traditional ASR systems, allowing for easier customization to specific tasks or domains, and when benchmarked with a HMM architecture, end-to-end models showed better results (Wang, Wang, & Lw, 2019). They can be trained on small amounts of data and can adapt to new speakers or languages with minimal additional training. Some studies have reported promising results using end-to-end ASR models in ATC settings, demonstrating that these models can achieve high levels of accuracy and can potentially outperform traditional ASR systems (Lin, Yang, Li, et al., 2021). While several end-to-end models exist, such as the combination of Convolutional Neural Networks (CNN) and RNN (Lin, Yang, Guo, & Fan, 2021; Fan, Guo, Lin, Yang, & Zhang, 2021; Li, et al., 2019), the most common ones are listed in the following sections.

3.3.1 Connectionist temporal classification (CTC)

CTC is a popular approach for end-to-end ASR. It uses a RNN to map input speech features directly to output sequences of characters or words. CTC allows for variable-length input and output sequences, making it well-suited for ASR (Chan, Jaitly, Le, & Vinyals, 2016-May).

3.3.2 Attention-based models

Attention-based models are another popular approach for end-to-end ASR. They use an encoder-decoder framework, where the encoder maps input speech features to a fixed-length representation and the decoder generates the output sequence. Attention mechanisms are used to allow the decoder to focus on different parts of the input representation at each decoding step, improving the accuracy of the system.

3.3.3 Recurrent neural networks with transducer (RNN-T)

RNN-T combines an encoder-decoder framework with a transducer to directly map variable-length input speech features to variable-length output sequences of characters or words. Unlike other end-to-end models, RNN-T allows for joint training of the acoustic and language models, enabling the system to adapt to the specific characteristics of the speaker and the language. Additionally, RNN-T can handle online recognition, where the system produces output in real-time as the speech is being input and has been proven to overperform CTC methodologies in speech transcription tasks (Prabhavalkar, et al., 2017).

3.4 RNN-T models

3.4.1 Architecture

The RNN-T model is composed of three main components: an encoder network, a prediction network, and a joint network as depicted on Figure 2 (Kanishka Rao, Hasim Sak, & Rohit Prabhavalkar, 2017). The encoder network is responsible for mapping the input acoustic frames into a higher-level representation, which is typically a fixed-length vector that captures the relevant information in the input speech signal. This representation is then used by the prediction network to generate an output sequence, which is typically a sequence of characters or words.

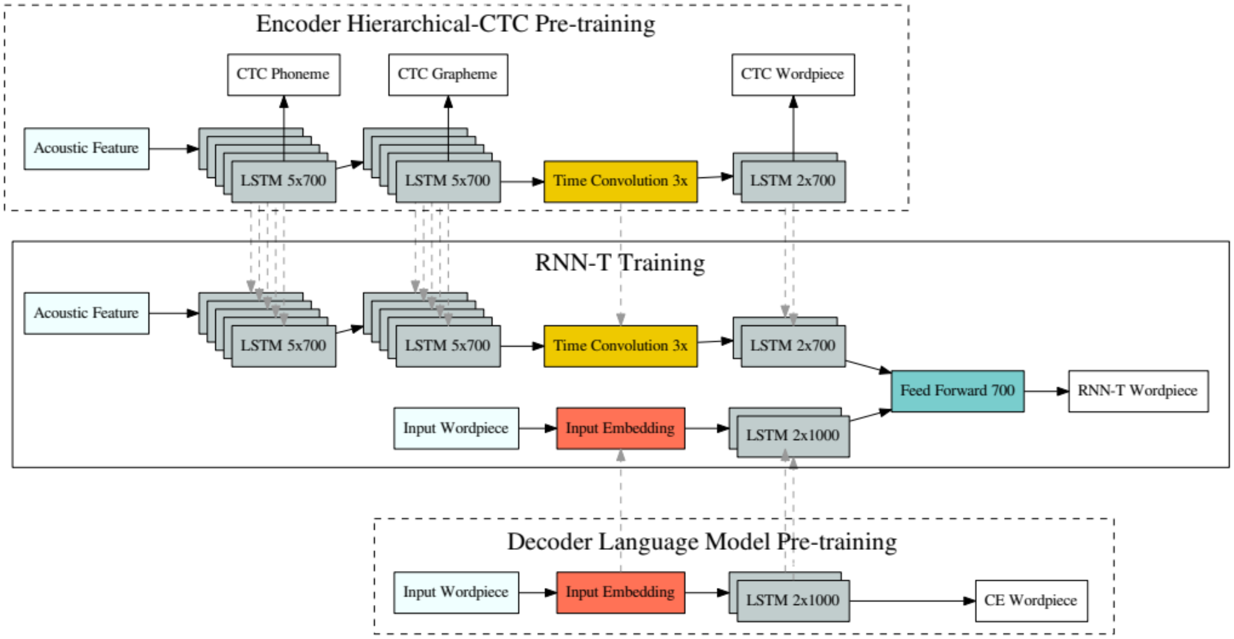

Figure 2. RNN-T architecture .

The prediction network corresponds to the decoder network and is conditioned on the history of previous predictions. It takes as input the output from the previous time step and generates the output for the current time step. The joint network combines the information from the encoder and prediction networks to generate the final output sequence.

3.4.2 Model training:

To train a RNN-T model, various stages are involved. First, the encoder network is pre-trained as a hierarchical-CTC network that can simultaneously predict phonemes, graphemes, and word pieces at different LSTM layers, such as 5, 10, and 12 layers. In addition, a time convolutional layer is used to reduce the encoder time sequence length by a factor of three.

The decoder network, on the other hand, is trained as an LSTM language model that can predict word pieces and is optimized with a cross-entropy loss. After training the encoder and decoder networks separately, the weights of the two pre-trained models are initialized and then combined to form the complete RNN-T network. The dashed lines in the diagram indicate the transfer of weights from the pre-trained models to the RNN-T network.

Finally, the complete RNN-T network is trained using the RNN-T loss function. The RNN-T loss is a joint optimization objective that optimizes both the acoustic and language models simultaneously (Kanishka Rao, Hasim Sak, & Rohit Prabhavalkar, 2017).

3.5 Deep speech

The Deep Speech model focuses on improving the accuracy and efficiency of speech recognition systems using DL techniques (Hannun, et al., 2014). The authors addressed some of the major challenges in traditional speech recognition systems, such as handling noisy and variable speech input, recognizing different accents and dialects, and scaling the model to handle large amounts of data. The team's contributions have significantly improved the accuracy and accessibility of speech recognition technology, opening new possibilities for human-computer interaction and communication.

The Deep Speech model uses a variant of RNN called a Bidirectional Recurrent Neural Network (BRNN) (Schusater & Paliwal, 1997), which combines information from both forward and backward sequences of the input speech signal. This allows the model to capture contextual information from both past and future frames of the speech signal, improving its ability to recognize speech accurately. We report an illustration of the architecture in Figure 3. Authors also introduced several techniques to improve the training process to enable the model to learn more efficiently from a diverse range of speech data.

When applied to an ATC environment, Deep Speech showed promising results (Kleinert, et al.). In addition of using the word error rate as an accuracy metric, this paper also measures the callsign recognition rate. This highlights the importance of semantic extraction for applications of ASR in an ATC environment.

3.5.1 Architecture & Computational efficiency

3.5.1.1 Architecture

Figure 3 (Hannun, et al., 2014) shows the model architecture which consists of five layers of hidden units.

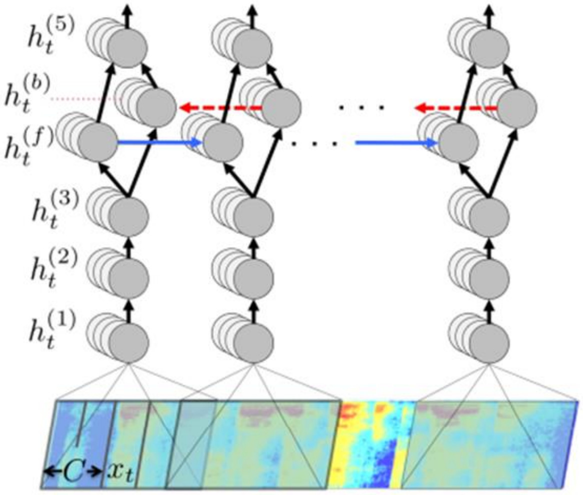

Figure 3. Deep Speech architecture.

The first three layers are non-recurrent layers. The hidden units are computed as follows.

Non-recurrent layer equations: Hidden unit computation:

$$h_t^{(l)} = g\left(W^{(l)} h_t^{(l-1)} + b^{(l)}\right) \tag{1}$$

where $g(z) = \min\{\max\{0, z\}, 20\}$ is the clipped rectified-linear (ReLu) activation function, t represents the time step, $W^{(l)}$, $b^{(l)}$ are the weight matrix and bias parameters for layer l and $h_t^{(l)}$ is the unit in the layer l at the time step t .

The fourth layer is a bi-directional recurrent layer composed of two sets of hidden units:

1. The first is built with forward recurrence:

Bi-directional recurrent layer equations: Exploring forward recurrence.

$$h_t^{(f)} = g\left(W^{(4)}h_t^{(3)} + W_r^{(f)}h_{t-1}^{(f)} + b^{(4)}\right) \tag{2}$$

where $h_t^{(f)}$ is the unit in the forward layer f at time step t .

2. The second is built with backward recurrence:

Bi-directional recurrent layer equations: Exploring backward recurrence:

$$h_t^{(b)} = g\left(W^{(4)}h_t^{(3)} + W_r^{(b)}h_{t+1}^{(b)} + b^{(4)}\right) \tag{3}$$

where $h_t^{(b)}$ is the unit in the backward layer b at time step t .

The fifth layer takes both the outputs of the forward and backward units as inputs:

Fifth layer equations: Combined outputs:

$$h_t^{(5)} = g\left(W^{(5)}h_t^{(4)} + b^{(5)}\right) \tag{4}$$

$where\, h_t^{(4)} = h_t^{(f)} + h_t^{(b)}$

The output layer is a standard $soft$ $\max(\cdot)$ function that yields the predicted character probabilities for each time slice t and character k in the alphabet:

Softmax output layer equation: Character probabilities:

$$h_{t,k}^{(6)} \equiv P(c_t = k|x) = \frac{\exp\left(W_k^{(6)} h_t^{(5)} + b_k^{(6)}\right)}{\sum_j \exp\left(W_j^{(6)} h_t^{(5)} + b_j^{(6)}\right)} \qquad (5)$$

where $W_k^{(6)}$ and $b_k^{(6)}$denote the $k'th$ column of the weight matrix and $k'th$ bias, respectively, and $c_t$is the predicted character at the time step t .

The CTC loss is used to measure the error in prediction during training, and Nesterov's Accelerated Gradient method is used to compute the gradient with respect to all the model parameters via back-propagation.

3.5.1.2 Computational efficiency

One of this research's contributions is implementing an effective training process, which has resulted in an accelerated performance of NNs. This methodology entails utilizing specialized networks that leverage high-speed computer operations and training them using multiple graphics processing units (GPUs) in parallel. The data and neural network model were partitioned into smaller subsets, which were then processed simultaneously on different GPUs. To speed up the procedure even more, instances of similar length were grouped together.

This method allowed the authors to process 2300 hours of data in just a few hours.

3.5.2 Dataset description

Table 1 shows the data used to train and evaluate Deep Speech model which is a collection of speech recordings from a variety of sources, including:

- WSJ (Wall Street Journal) is a corpus of reading speech recordings from the Wall Street Journal. It consists of about 80 hours of speech from 280 speakers and is commonly used for training and testing speech recognition models.
- SWITCHBOARD is a corpus of conversational speech recordings collected over the telephone network. It consists of about 2,400 two-sided conversations between two strangers and is commonly used for training and evaluating speech recognition models.
- FISHER is another corpus of conversational speech recordings, collected in a similar way to SWITCHBOARD. It consists of about 2,000 conversations between native speakers of American English and non-native speakers and is often used for evaluating speech recognition models in challenging conditions.

- Baidu is a dataset collected by Baidu Research, consisting of many spoken sentences in Mandarin Chinese. It contains both clean and noisy recordings and is often used for training and evaluating speech recognition models in Mandarin Chinese.

Table 1. The used benchmark datasets

Dataset	**Type**	**Hours**	**Speakers**[i]
WSJ	read	80	280
Switchboard	conversational	300	4000
Fisher	conversational	2000	23000
Baidu	read	5000	9600

The authors' objective is to enhance the functionality of existing systems that fail to perform efficiently in noisy surroundings. However, acquiring labeled data from such environments is challenging. To tackle this issue, they devised an alternative approach for generating data, which involved combining 100 noisy and 100 noise-free utterances from 10 speakers with SNR between 2 and 6dB. This data was created to enhance performance in noisy settings.

The Lombard effect (List, 1993) occurs when speakers change their voice's intonation or pitch to overcome background noise. Recorded speech datasets, which are frequently gathered in quiet environments, do not, however, capture this effect. To overcome this limitation the authors played loud background noise through the headphones while the person was speaking, thus ensuring that the Lombard effect would be present in the recordings.

3.5.3 Experiments and results

3.5.3.1 First experiment: Conversational speech: Switchboard Hub5'00 (full)

The authors tested their system on a difficult dataset, Hub5'00, which has both easy (SWB) and hard (CH) instances and reported the word error rate for the full set. They trained their model on a 300-hour dataset of Switchboard speech on a larger 2300-hour dataset combining Switchboard and Fisher speech. They computed spectrograms of 80 linearly spaced log filter banks and applied speaker adaptation by normalizing the spectral features for each speaker. They used a 4-gram language model with a 30,000-word vocabulary for decoding. The Deep Speech SWB model has 5 hidden layers with 2048 neurons, trained only on the 300-hour Switchboard dataset. The Deep Speech SWB + FSH model is an ensemble of 4 RNNs, each with 5 hidden layers of

2304 neurons, trained on the full 2300-hour dataset. Both models were trained on inputs of +/- 9 frames of context.

In Table 2, we can see a comparison of the proposed models with other of the state-of-the-art models. They use word error rate (WER) as the evaluation metric. Vesel et al. (2013) used a hybrid DNN-HMM system with a sequence-based loss function and had the best previously published result on the Hub5'00 test set. However, when they trained their Deep Speech system on the combined 2300 hours of data, it improved the performance by 2.4% absolute WER. DNN-HMM FSH, developed by Maas et al. (2017), achieved a 19.9% WER when trained on the Fisher 2000-hour corpus using Kaldi, another open-source ASR software. This result shows that the proposed Deep Speech system is competitive with the best existing ASR systems when trained on a similar amount of data.

Table 2. Published error rates (%WER) on Switchboard dataset splits.

Model	**SWB**	**CH**	**Full**
Vesely et al. (GMM-HMM BMMI) (Vessel, Ghoshal, Burget, & Povey, 2013)	18.6	33.0	25.8
Vesely et al. (DNN-HMM sMBR) (Vessel, Ghoshal, Burget, & Povey, 2013)	12.6	24.1	18.4
Maas et al. (DNN-HMM SWB) (Maas, et al., 2017)	14.6	26.3	20.5
Maas et al. (DNN-HMM FSH) (Maas, et al., 2017)	16.0	23.7	19.9
Seide et al. (CD-DNN) (Seide, Li, Chen, & Yu, 2011)	16.1	n/a	n/a
Kingbury et al. (DNN-HMM sMBR HF) (Kingsbury, Sainath, Soltau, Watson, & Heights, 2012)	13.3	n/a	n/a
Sainath et al. (CNN-HMM) (Sainath, et al., 2013)	11.5	n/a	n/a
Soltau et al. (MLP/CNN+I-Vector) (Watson, Heights, Soltau, Saon, & Sainath, 2023)	**10.4**	n/a	n/a
Seep Speech SWB	20.0	31.8	25.9
Deep Speech SWB + FSH	12.6	19.3	16.0

3.5.3.2 Second experiment: Noisy speech

The second set of experiments involved testing the performance of the proposed speech recognition models in noisy environments. They used 100 noisy and 100 noise-free recordings from 10 speakers and created the noise using various environments like a crowded cafeteria, a restaurant, and driving in the rain. They trained their model using more than 7000 hours of data and used an ensemble of 6 networks. They compared their model to other commercial speech systems and found that their system performs better in noisy environments. They also trained two RNNs, one on raw data and the other with added noise and found that the one trained with noise performed better in noisy environments.

The results in Table 3 indicate that the Deep Speech model outperformed the commercial systems in noisy environments, and the noise-trained model achieved a 6.56% absolute on the clean recordings, 19.06% on the Noisy data, and 11.85% on the combined version.

Table 3. Results (%WER) for 5 systems.

System	**Clean (94)**	**Noisy (82)**	**Combined (176)**
Apple Dictation	14.24	43.76	26.73
Bing Speech	11.73	36.12	22.05
Google API	6.64	30.47	16.72
wit.ai	7.94	35.06	19.41
Deep Speech	**6.56**	**19.06**	**11.85**

3.6 Transformers

3.6.1 Introduction

Transformers have emerged as a dominant force in the realms of audio and NLP tasks. With their novel architecture and self-attention mechanisms, Transformers have achieved state-of-the-art performance, surpassing previous approaches in understanding and generating human language. In NLP, they excel in machine translation, sentiment analysis, question answering, and text summarization. In speech tasks, they enabled more accurate transcriptions and natural-sounding speech generation. Transformers represent a remarkable advancement that has reshaped the landscape of language-based AI applications.

In the context of ASR, Transformers have been proven to outperform techniques using RNNs such as RNN-T or LSTM cells due to their parallelization capabilities (Zeyer, Bahar, Irie, Schluter, & Ney, 2019; Vaswani, et al., 2017).

3.6.2 Components

3.6.2.1 Semantic encoding

Semantic encoding constitutes a fundamental step in NLP, serving to enhance machines' ability to comprehend the significance of textual data. It involves transforming natural language text into a numerical representation that captures the context and semantics of the text.

The Transformer architecture places considerable importance on semantic encoding, which serves as a critical step in transforming the input sequence into a series of compact and dense vector representations. These representations encapsulate the significance and contextual information of each token, facilitating the model's ability to comprehend the interdependence between different components of the input sequence. Figure 4 demonstrates how the semantic encoding step in the Transformer architecture effectively preserves the semantic relationships within the input text.

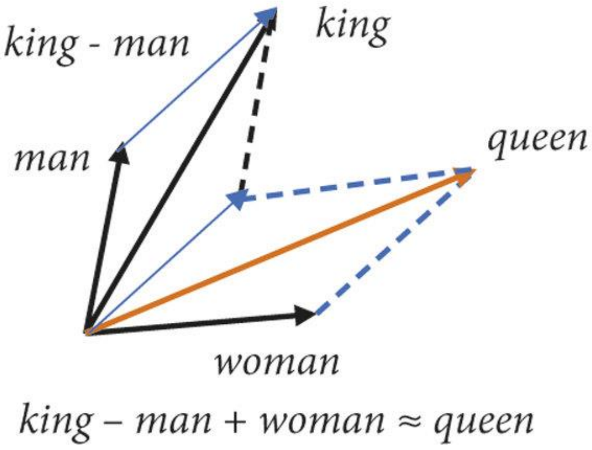

Figure 4. Sample semantic relationships

3.6.2.2 Positional encoding

Positional encoding is important in the Transformer architecture for NLP. It helps the model understand the order of input tokens in a sequence which is important for understanding the meaning of the text. Positional encoding assigns a unique vector to each token that represents its position in the sequence.

The Transformer's positional encoding that the designers proposed is to use a periodically varying function based on sine and cosine here are the formulas:

Sine-cosine positional encoding for the transformer model:

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right), for\ i = 0,\ 1,\ 2, \dots, \frac{1}{2}d_{model} - 1 \tag{6}$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right), for\ i = 0,\ 1,\ 2, \dots, \frac{1}{2}d_{model} - 1 \tag{7}$$

where ‘pos ‘ is the position (time-step) in the sequence of input words, ‘i ‘ is the position along the embedding vector dimension and ‘d_{model}‘represents the dimension of the embedding vectors.

The size of the positional encoding vector is identical to that of the input embedding vector, and the former is added to the latter to obtain the final input representation. The incorporation of positional encoding into the Transformer architecture is instrumental in enabling the model to differentiate between tokens that may possess identical representations but occupy different positions within a sequence. Consequently, this feature plays a pivotal role in facilitating the Transformer's ability to comprehend the sequential order of input tokens, thereby allowing for the generation of highly accurate output sequences.

3.6.2.3 Self-attention mechanism

The self-attention mechanism is one of the key innovations of the Transformer architecture, particularly in the context of NLP and audio tasks. This mechanism enables the models to capture long-range dependencies between tokens in the input and output sequences more effectively than traditional models that rely on RNNs or CNNs.

Figure 5 shows the three distinct types of attention mechanisms that are used in the Transformer to improve its performance:

1. **Encoder-Decoder Attention**: Attention between the input sequence and the output sequence.
2. **Self-attention in the input sequence:** Attention between all the words in the input sequence.
3. **Self-attention in the output sequence**: The purpose of the masking in the self-attention mechanism of the decoder is to ensure that each token in the output sequence attends only

to the previously generated tokens, and not to any future tokens. This prevents the model from "cheating" by looking ahead at tokens that it has not generated yet. This is done by masking the words that occur after it for each step. So, for step 1, only the first word of the output sequence is NOT masked, for step 2, the first two words are NOT masked and so on.

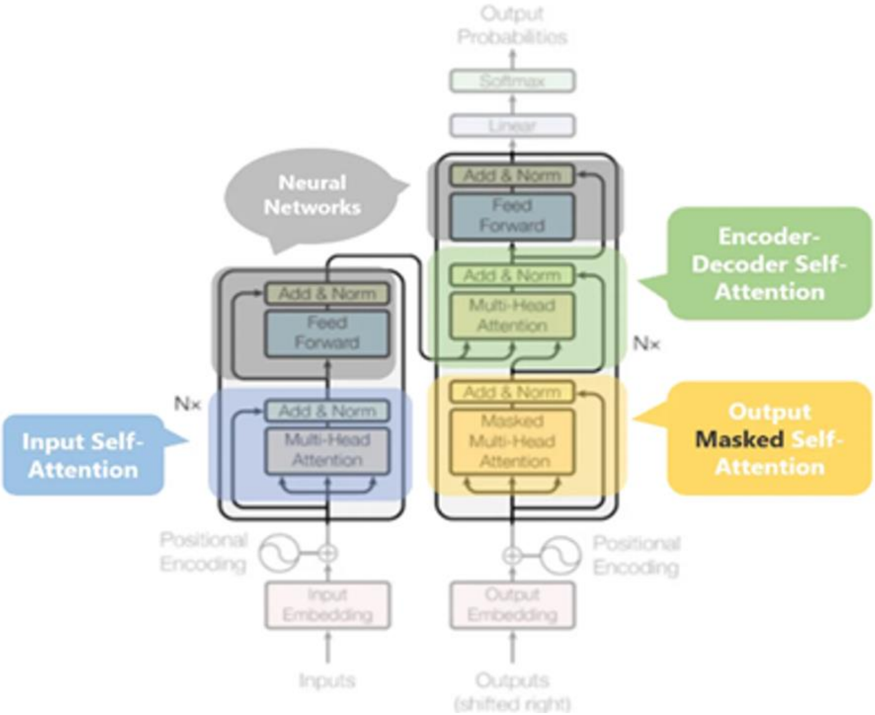

Figure 5. Attention mechanisms in Transformer

The self-attention mechanism involves computing a weighted average of the input vectors, with emphasis placed on the most crucial vectors.

Weighted vector aggregation: Self-attention mechanism:

$$z_i \; = \sum_{k=0}^{n} w_{ik} \, x_k \tag{7}$$

where ‘k ’ indexes over the complete sequence of input vectors, ‘n ‘ is the number of input vectors and w_{ik} is the attention weight derived from input vectors computed as follows:

Attention weights computation:

$$w_{ik} = \mathit{soft}\max(x_i^T x_k) \tag{8}$$

where 'x_i' is at the same position as 'z_i', 'T ' is the transpose operation and $\mathit{soft}\max(\cdot)$ is used to map the values between 0 and 1

In the basic self-attention operation, each input vector 'x_i' is used in three distinct roles.

Multifaceted roles: Self-attention output equation with three roles for input vectors:

$$z_i = \sum_j \mathit{soft}\max\left(x_i^T x_j\right) x_j \tag{9}$$

- These roles are called query, key, and value.
- The self-attention mechanism uses learnable parameters to derive three different vectors for the roles of the query, key, and value. These vectors are obtained through a linear transformation of the original input vectors.

Learnable Transformations: Query, Key, and Value Vectors in Self-Attention Mechanism.

$$\begin{aligned} q_i &= W^q x_i, \\ k_i &= W^k x_i, \\ v_i &= W^v x_i, \end{aligned} \tag{10}$$

where 'W^q', 'W^k' and 'W^v' are learnable weight matrices.

This feature enables the model to focus on the most salient aspects of the input sequence during both encoding and decoding, thereby enhancing its ability to comprehend the intricate relationships between various segments of the sequence.

3.6.3 Multi-head attention

Multi-head attention constitutes a critical component within the Transformer architecture, playing a pivotal role in enabling the model to concurrently attend to multiple segments of the input sequence. This improves the model's ability to capture more complex patterns and relationships between words, which can be helpful in both NLP and audio tasks. After processing each constituent part of the input sequence, the model amalgamates all the resulting information by aggregating the outputs from each component, as illustrated in Figure 6. The model subsequently utilizes mathematical techniques to optimize the aggregated data, enhancing its usefulness in generating highly accurate output sequences. This optimized data is then passed through a neural network, allowing the Transformer model to produce even more refined output sequences.

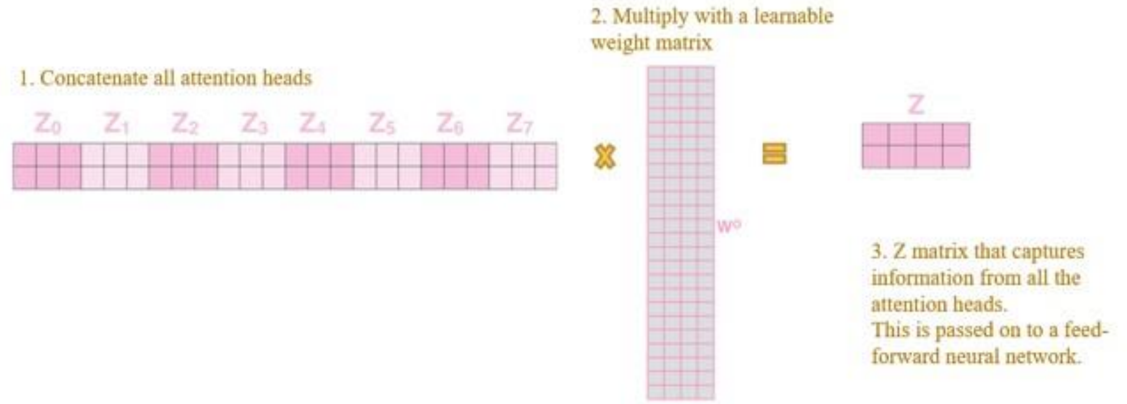
1. Concatenate all attention heads
Z_0 Z_1 Z_2 Z_3 Z_4 Z_5 Z_6 Z_7
2. Multiply with a learnable weight matrix
W^O
Z
3. Z matrix that captures information from all the attention heads.
This is passed on to a feed-forward neural network.

Figure 6: Multi-head attention mechanism

3.6.4 Transformer architecture

As we can see in Figure 7, the Transformer architecture is composed of an encoder and a decoder.

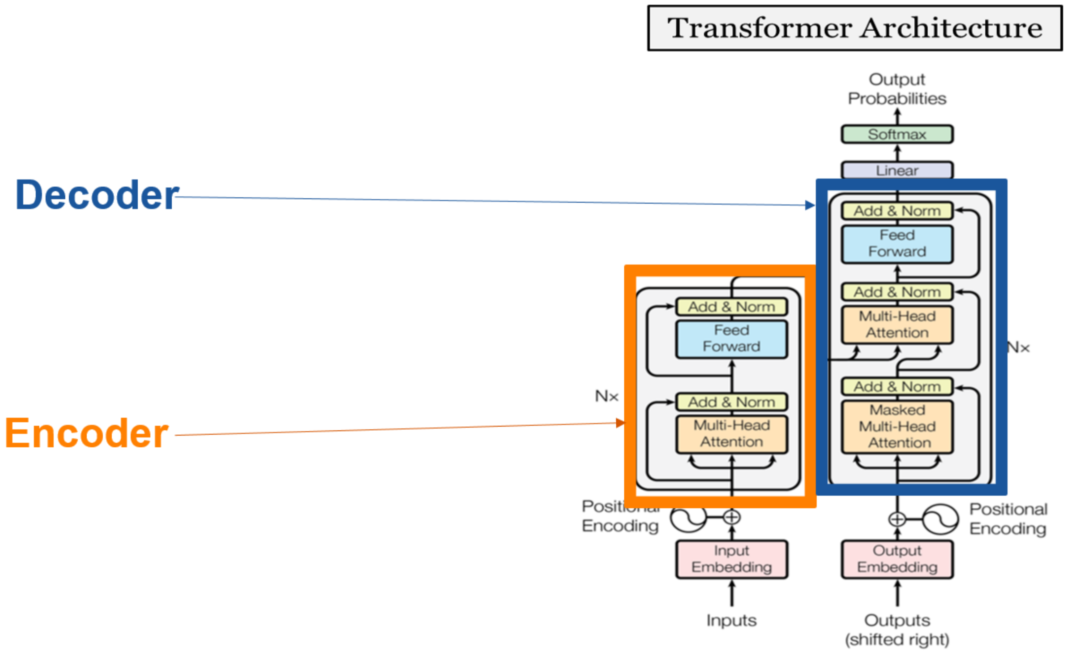

Figure 7. Transformer architecture

The Transformer architecture is a neural network model that has revolutionized the field of audio and NLP. It consists of an encoder and a decoder, which work together to process and generate sequences of text. In the next sections, we will delve deeper into the workings of the encoder and decoder components of the Transformer architecture and examine how they contribute to the overall success of the model in various audio and NLP tasks.

3.6.4.1 Encoder

The first step in the Transformer model is to convert the input sequence into a set of vector representations that encapsulate the semantic meaning of each word. These vectors are then augmented with positional encoding to indicate the position of each word in the sequence. In Figure 8, we observe that the vectors are subsequently processed through a multi-head attention layer, which generates attention scores for all pairs of positions in the sequence. This mechanism enables the model to focus on the most relevant aspects of the sequence. Finally, the output of the self-attention sub-layer is passed through a feed-forward neural network to capture intricate relationships between different parts of the input sequence.

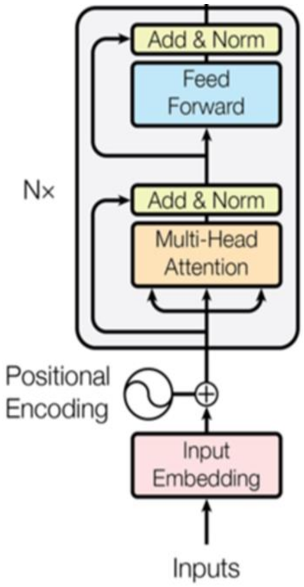

Figure 8. The encoder of transformer

The Encoder is designed to be highly parallelizable, allowing it to process input sequences efficiently and in parallel. The outputs of the Encoder are a set of dense vector representations. These vector representations are then fed into the Decoder component of the Transformer architecture.

3.6.4.2 Decoder

In the decoding stage of the Transformer model, the output sequence generated thus far is utilized as the input to the decoder. The embedding of each output token is augmented with positional encoding, which is shown in Figure 9, and these resulting vectors are then passed through the first layer of the decoder. In each subsequent layer, the masked multi-head attention mechanism computes attention scores between every pair of positions in the output sequence up to the current time step, ensuring that the model only attends to previously generated tokens.

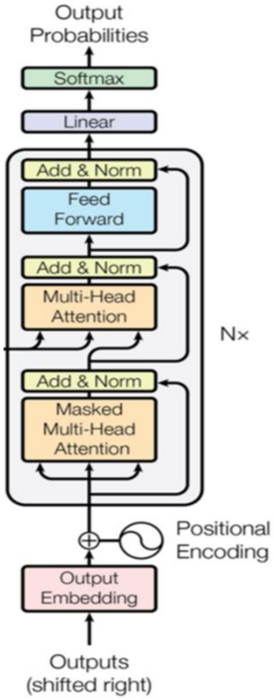

Figure 9. The decoder of transformer

This mechanism boosts the model's ability to generate precise output sequences by leveraging previously generated tokens to guide subsequent predictions. The masked self-attention sub-layer's outputs are then passed through a multi-head attention mechanism that attends to the encoded input sequence produced by the Encoder. This mechanism calculates attention scores between every position in the output sequence and every position in the encoded input sequence, allowing the model to focus on the relevant portions of the encoded input sequence at each step of the decoding process. Finally, the outputs of the multi-head attention sub-layer are passed through a feed-forward neural network, which employs a non-linear transformation to enable the model to capture more intricate relationships between the input and output sequences.

The Decoder is also designed to be highly parallelizable, allowing it to generate the output sequence efficiently and in parallel. The final output of the Decoder is the next predicted token.

3.6.5 Use cases

Transformers have been widely used for ASR in recent years. For example, Mohamed et. al. (2019) present a new approach for ASR that combines Transformers and CNNs to capture both local and global contextual information from speech data. The proposed model, Convolutional Context Transformer, uses CNNs to extract local features and Transformers to model global context dependencies. The model is trained on the LibriSpeech dataset and evaluated on the test-clean and test-other subsets, achieving state-of-the-art results.

In the context of Transformers, there exist two main categories depending on how the model generates an output. Autoregressive models generate output one element at a time, conditioned on previously generated elements, while non-autoregressive models generate output all at once, without considering any previously generated elements. In the context of ASR, autoregressive models have been widely used and achieve state-of-the-art results, but they can be computationally expensive and slow due to their sequential nature. Non-autoregressive models, on the other hand, are faster and can generate output in parallel, but they often sacrifice some accuracy for speed. Song et al. proposed an ASR model that combines a non-autoregressive Transformer encoder with a CTC-enhanced decoder input. The model was trained jointly with a hybrid CTC/attention-based loss function, which allows for faster training and better convergence than previous non-autoregressive models. Moreover, the proposed approach achieved a 50 times faster decoding speed compared to a strong autoregressive model (Song, et al., 2021).

Although it was shown that transformer-based models are a strong alternative to RNN end-to-end models, these approaches require the entire input sequence to compute the self-attention, making them computationally expensive. To mitigate this obstacle, another study proposes a new approach to enhance the performance of Transformer-based ASR models by introducing a new block-level processing method called Contextual Block Processing (CBP). The proposed method aims to incorporate contextual information of input blocks into the Transformer encoder and decoder layers to improve the model's ability to capture long-term dependencies (Tsunoo, Kashiwagi, Kumakura, & Watanabe, 2019).

3.6.6 Conclusion

In conclusion, Transformers have had a significant impact on audio and NLP fields, and their innovative architecture has been successfully integrated into various algorithms for a range of tasks such as sentiment analysis, machine translation, and speech-to-text. With their ability to capture long-range dependencies, attend to different parts of the input sequence simultaneously, and effectively encode positional information, Transformers have shown great promise in advancing audio and NLP and improving the accuracy and performance of language-based applications. Transformers are able to understand complex relationships between different tokens of input and output sequences thanks to the self-attention mechanism and the multi-head attention mechanism. Additionally, the highly parallelizable Transformers architecture has shown a faster and more efficient way to produce predictions compared to traditional methods.

3.6.7 Key takeaways

4. Transformers have become the de facto standard for audio and NLP tasks.
5. Semantic encoding, positional encoding, self-attention mechanism, and multi-head attention mechanism are the key components of the Transformers architecture.
6. The ability of Transformers to learn contextual representations of words has significantly improved the performance of language models.
7. Transformers have proven to be highly parallelizable, enabling the generation of predictions in a faster and more efficient way.
8. The pre-trained language models based on Transformers, such as BERT (Kenton, Kristina, & Devlin, 1953) and GPT (Redford & Salimans, Redford & Salimans, n.d.), have achieved state-of-the-art performance on various natural language processing tasks.
9. The Whisper model (Radford, et al., 2022), a pre-trained audio model inspired by Transformers, has demonstrated remarkable performance in a wide range of audio processing tasks.
10. Transformers have enabled significant progress in the development of conversational agents and chatbots.

4 References

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016-May). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *ICASSP, IEEE International Converence on Acoustics, Speech and Signal Processing - Proceedings*, (pp. 4960-4964). Retrieved from https://doi.org/10.1109/ICASSP.2016.7472621

Chen, S., Kopald, H. D., Chong, R. S., Levonian, Z., Wei, Y., & White, K. (2016, September). Methods for expanding speech recognition applications for early resolution of surface safety events.

Dahl, G. E., Member, S., Deng, L., & Acero, A. (2011, April). *Context-dependent pre-trained deep neural netwroks for large-vocabulary speech recogntion.* doi:10.1109/TASL.2011.2134090

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: features, classification schemes, and databases. *Pattern Recognition, 44*(3), pp. 572-587. Retrieved from https://doi.org/10.1016/j.patcog.2010.09.020

Fan, P., Guo, D., Lin, Y., Yang, B., & Zhang, J. (2021). *Speech recogntion for air traffic control via feature learning and end-to-end training.* Retrieved from http;//arxiv.org/abs/2111.02654

Ferreiros, J., Pardo, J. M., De Córdoba, R., MacIas-Guarasa, J., Montero, J. M., Fernández, F., . . . González, G. (2012). A speech interface for air traffic control terminals. *Aerospace Science and Technology, 21*(1), pp. 7-15. Retrieved from https://doi.org/101016/j.ast.2011.05.002

Geacăr, C. M. (2010). Reducing pilot/ATC communication errors using voice recognition. In I. 2010 (Ed.), *27th Congress of the International Council of the Aeronautical Sciences 2010, 6,* pp. 4866-4872.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., . . . Ng, A. Y. (2014). *ARXIV.org.* Retrieved from Deep speech: scaling up end-to-end speech recognition.: http://arxiv.org/abs/1412.5567

Helmke, H., Ohneiser, O., Muhlhausen, T., & Wies, M. (2016). Reducing controller workload wtih automatic speech recognition. *AIAA/IEEE Digital Avionics Systems Conference - Preceedings*, (pp. 1-10). Retrieved December 2016, from https://doi.org/10.1109/DASC.2016.7778024

Henriques, R. (2009). Initial Investigartion of speech recognition capabilities for FAA airspace security operations.

Kanishka Rao, Hasim Sak, & Rohit Prabhavalkar. (2017). Exploring architectures, data and units for streaming end-to-end speech recognition with RNN-transducer. *Conference: 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, (pp. 193-199). doi:10.1109/ASRU.2017.8268935

Kenton, M. C., Kristina, L., & Devlin, J. (1953). BERT: Pre-training of deep bidirectional transformers for language understanding. Mlm.

Kingsbury, B., Sainath, T. N., Soltau, H., Watson, I. B., & Heights, Y. (2012). Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization. *Interspeech 2012: ISCA's 13th Annual Conference*, (pp. 10-13). Retrieved from http://www.isca-speech.org/archive

Kleinert, M., Venkatarathinum, N., Helmke, H., Ohneiser, O., Strake, M., & Fingscheidt, T. (n.d.). Easy adaptation of speedh recognition to different air traffic control environments using the DeepSpeech engine. *11th SESAR Innovation Days*, 1-8. Retrieved January 2021

Kocour, M., Vesely, K., Gomez, J. Z., Szoke, I., Chernocky, J., "Honza" Klakow, D., & Motlicek, P. (2021). Boosting of contextual information in asr for air-traffic call-sign recognition. *Proceedings fo the annual conference of the International Speech Communication Association, INTERSPEECH., 1*, pp. 256-260. Retrieved 2021, from https://doi.org/10.21437.Interspeech.2021-1619

Kopald, H. (2017, September). Using speech recognition to identify the initiator of the missed approach. 1-33.

Kopald, H., & Chen, S. (2019). Real-time applications of automatic speech recognition technology for the National Airspace System Center for Advanced Aviation System Development.

Kopald, H., Chong, R. S., & Shepley, K. (2018). *Speech recognition for real-time surface safety applications.*

Li, J., Lavrukhin, V., Ginsburg, B., Leary, R., Kuchaiev, O., Cohen, J. M., . . . Gadde, R. T. (2019). Jasper: An end-to-end convolutional neural acoustic model. *Preceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, (pp. 71-75). Retrieved September 2019, from https://doi.org/10.21437/Interspeech.2019-1819

Lin, Y., Guo, D., Zhang, J., Chen, Z., & Yang, B. (2021). A unitified framework for multilingual speech recognition in air traffic control systems. *IEEE Transactions on neural networks and learning systems, 32(8),* pp. 3608-3620. Retrieved from https://doi.org/10.1109/TNNLS.2020.3015830

Lin, Y., Tan, X., Yang, B., Yang, K., Zhang, J., & Yu, J. (2019). Real-time controlling dynamics senseing in air traffic system. *19*(3). Retrieved from https://doi.org/10.3390/s19030679

Lin, Y., Yang, B., Guo, D., & Fan, P. (2021). Towards multilingual end-to-end speech recognition for air traffic control. *15*(9), pp. 1203-1214. Retrieved from https://doilorg/10.1049/itr2.12094

List, S. C. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers.

Maas, A. L., Qi, P., Xie, Z., Hannun, A. Y., Lingerich, C. T., Jurafsky, D., & Ng, A. Y. (2017). Building DNN acoustic models for large vocabulary speech recognition. *Computer Speech & Language, 41*, 195-213.

Maas, A., Le, Q. V., O'neil, T. M., Vinyals, O., Nguyen, P., & Ng, A. Y. (2012). Recurrent neural networks for noise reduction in robust ASR.

McGuire, R., & Feerrar, W. (2014). *Voice data.* The MITRE Corporation.

Mohamed, A., Okhonko, D., & Zettlemoyer, L. (2019, April 26). Transformers with convolutional context for ASR. *Computer Science > Computation and Language,* pp. 1-5. Retrieved from http://arxiv.org/abs/1904.11660

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *7*, 19143-19165. Retrieved from https://doi.org/10.1109/ACCESS.2019.2896880

Povey, D. (2004). Discriminative training for large vocabulary speech recognition.

Prabhavalkar, R., Rao, K., Sainath, T. N., Li, B., Johnson, L., & Jaitly, N. (2017). A compression of sequence-to-sequence models for speech recognition. *Proceedings of the annual conference of the International Speech Communication Association.* (pp. 939-943). INTERSPEECH. Retrieved from https://doi.org/10.21437/Interspeech.2017-233

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. doi:10.48550/arXiv.2212.04356

Reddy, D. R. (1976). Speech recognition by machine: a review. *Preceedings of the IEEE, 64(4), 64(4)*, pp. 501-531. Retrieved from https://doi.org/10.1109/PROC.1976.10158

Redford, A., & Salimans, T. (n.d.). *Improving language understanding by generative pre-training.* Retrieved from https://www.semanticscholar.org

Renkens, V. (. (2017). Deep learning in automatic speech recognition: What is automatic speech recognition. *ASR, 37*(1). Retrieved from https://doi.org/10.2991/978-2-494069-51-0

Rodríguez, E., García-Crespo, Á., & García, F. (1997). *Speech/speaker recognition using a HMM/GMM hybrid model.* Retrieved from Springer.com: https://doi.orf/10.1007/bfb0016000

Sainath, T. N., Kingsbury, B., Mohamed, A., Dahl, G. E., Saon, G., Soltau, H., . . . Heights, Y. (2013, September). *Improvements to deep convolutional neural networks for LVCSR.* doi:10.1109/ASRU.2013.6707749

Schusater, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *Computer Science, 45*(11), pp. 2673-2681. doi:DOI:10.1109/78.650093Corpus ID: 18375389

Seide, F., Li, G., Chen, X., & Yu, D. (2011). Feature engineeromg om context-dependent deep neural networks for conversational speech transcription. *2011 IEEE Workshop on Automatic Speech Recognition & Understanding,* (pp. 24-29). doi:10.1109/ASRU.2011.6163899

Song, X., Wu, Z., Huang, Y., Weng, C., Su, D., & Meng, H. (2021, April 16). Non-autoregressive transformerASr with CTC-enhanced decoder input. 5894-5898. Hng Kong, Shenzhen, China: Tsinghua Shenzhen INternational Graduat School, Tsinghua University, Shenzhen The Chinese University of Hong Kong. Jpmg Lpmg Temcemt AO :ab. doi:10.48550/arXiv.2010.15025

Tarakan, R. (2012). *Late or missing landing clearance detection and notification system description.*

Tsunoo, E., Kashiwagi, Y., Kumakura, T., & Watanabe, S. (2019). Transformer ASR with contextual block processing. *2019 IEEE Automatic speech recognition and understanding workshop.,* (pp. 427-433). Retrieved from https://doi.org/10.1109/ASRU46091.2019.9003749

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017, December). Attention is all you need. *Advances in Neural Information Processing Systems.*, pp. 5999-6009. doi:10.48550/arXiv.1706.03762

Vessel, K., Ghoshal, A., Burget, L., & Povey, D. (2013, August). Sequence-discriminative training of deep neural networks. (pp. 2345-2349). Interspeech 2013.

Vinalys, O., & Ravuri, S. V. (2011). Comparing mulltilayer perceptron to deep belief network tandem features for robust ASR. *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings.*, (pp. 4596-4599). Retrieved from https://doi.org/10.1109.ICASS/2011.5947378

Wang, D., Wang, X., & Lw, S. (2019). An overview of end-to-end automatic speech recognition. *Symmetry, 11*(8), 1-26. Retrieved from https://doi.org.10.3390/sym11081018

Watson, I. B., Heights, Y., Soltau, H., Saon, G., & Sainath, T. N. (2023). Join training of convolutional and non-convolutional neural networks. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 5572-5576). doi:10.1109/ICASSP.2014.6854669

Zeyer, A., Bahar, P., Irie, K., Schluter, R., & Ney, H. (2019). A comparison of transformer and LSTM encoder decoder models for ASR. *2019 IEEE Automatic Speech Recognition and Understanding Workshop. AsRU 2019 - Proceedings.*, (pp. 8-15). Retrieved from https>//doi.org/10.1109/ASRU46091.2019.9004025

A MITRE STT reports summarized

Table A- 1. Voice communications error detection & notification: an initial feasibility assessment (2005)

Summary	**Pros**	**Cons**
• The MITRE Corporation conducted a feasibility assessment of a voice communications error detection system in September 2005. • The project aimed to improve aviation safety by reducing operational errors and deviations caused by controller-pilot voice communications errors. • Automated Speech Recognition (ASR) technology was used to detect potential errors by looking for mismatches between controller clearances and the associated pilot readback. • Two evaluations were performed to determine if the speech recognition accuracy rates for operational ATC voice communications were similar to those observed on the CTI project for ATC trainees in a laboratory environment. • Results indicated that tuning the ASR dictionary for ATC specific pronunciation and Digital Signal Processing to slow down the audio speed improved recognition accuracy scores.	• Automated Speech Recognition (ASR) technology can be used to detect voice communications errors in aviation safety • ASR technology can recognize 7110.65 phraseology and colloquialisms • Digital Signal Processing (DSP) and ASR tuning techniques can improve recognition accuracy • Accuracy scores are highest for certain key information, such as the flight level digits	• Initial results indicated low speech recognition accuracy confidence scores • Accuracy scores were lower for filler words • Further research is needed to determine what recognition accuracy is needed to support Concept of Operations

Table A- 2. Issues to be addressed during full analysis and performance measurement of a voice communications error detection system (2005)

Summary	Pros
• Developing a voice communications error detection and notification system to improve aviation safety by reducing Operational Errors and Operational Deviations caused by controller-pilot voice communications errors. • Investigating the feasibility of using COTS Automated Speech Recognition (ASR) technology to recognize controller speech. • Adapting the ASR system to controller and pilot acoustic models. • Developing recognition grammar and intent rules to determine the intent of the issued ATC clearance. • Calculating expected correct and incorrect pilot readback for a control instruction. • Analyzing the connection between speech recognition capability and categories of voice communications errors. • Developing capability for recognition of pilot readbacks. • Calculating missed alarm/false alarm curves for readback-hearback error detection.	• Can improve aviation safety by reducing Operational Errors & Operational Deviations caused by controller-pilot voice communications errors. • Can identify the intent of the issued ATC clearance. • Can determine if the Pilot readback is consistent with the intent of the issued ATC clearance. • Can determine whether confidence in match or mismatch result is sufficient to notify ATC controller/system. • Can use recognition grammar & intent rules developed by CAASD for Controller Training Initiative. • Can use vendor provided tools & CAASD developed tools with controller voice recordings to collect missed recognition & false recognition rates. • Can use pilot voice tapes.

Table A- 3. Voice communications error detection assessment of speech recognition system adaptation (2006)

Summary	**Pros**	**Cons**
- The text is an assessment of Voice Communications Error Detection and Notification for an aviation safety output. - OE Analysis results found that voice communication errors represented approximately 20% of errors, with 50% involving single digit or letter errors and 25-33% involving transposed digits or letters. - To focus the effort, the detection system was targeted at altitude errors, callsign errors and lack of pilot response. - Phonetic and acoustic model adaptations were developed to modify the ASR speech models for ATC phraseology. - Evaluations showed a 10% improvement	- Both Phonetic & Acoustic Model Adaptation can improve ASR recognition rates. - Analysis of enroute OEs identified the most common voice communications problems. - Evaluation environment includes Automated Speech Recognition Systems, Speech Recognition Grammar, and ATC Speech Database.	- COTS ASR systems may not have sufficiently high recognition rates to support a general purpose ATC voice communications error detection system. - Not suitable for direct comparison between phonetic and acoustic model adaptation studies. - Evaluations only on the controller side of the dialogue.

Table A- 4. Late or missing landing clearance detection and notification system description (2012)

Summary
• MITRE CAASD designed and implemented a Late or Missing Landing Clearance Detection and Notification System prototype. • The prototype utilizes automatic speech recognition, data fusion, and other technologies to make it a viable and practical capability in the busy tower cab environment. • This document captures and describes the key elements of the prototype and its demonstration in the lab, including design considerations, software architecture, algorithms, hardware components, site specific parameters, data and configurations, and simulation environment. • The system performed well in the simulated environment and is ready for field customization and site-specific enhancements. • Additional airport/tower specific adaptations are still necessary to optimize the system for controller vocabulary, pronunciation, and phraseology and voice switch audio characteristics as well as airport automation systems and runway layout.

Table A- 5. Voice data (2014)

Summary	**Pros**	**Cons**
• MITRE Corporation developed a Flight Analysis System (FAS) capability to enable researchers to access recorded aviation data. • It evolved the capability to merge surveillance data from various sources with other aviation data sources to create a Flight Story. • Pilot controller voice communications are included to provide tactical pilot and controller intent information. • A voice processing pipeline of nine steps has been developed to incorporate voice data into the FAS. • The steps were audio filtering/sampling/transcoding, audio segmentation, dynamic context retrieval, dynamic callsign grammar generation, automatic speech recognition for callsigns, automatic speech recognition for full text transcription, extraction of information from text, fusion with Threaded Tracks, and encoding for storage	• Can provide researchers with easy access to recorded aviation data. • Accessibility of voice communications data on a large scale. • Provides information critical to understanding of communication related errors. • Automatically generates files that can be loaded into the Flight Analysis System (FAS). • Utilizes automatic speech recognition technology. • Includes dynamic context retrieval, dynamic callsign grammar generation and extraction of information from text.	• Limitation of not including military and other sensitive flights. • Difficulty in processing lower-fidelity audio. • Inability to identify Flight Story associated with individual transmissions.

Table A- 6. Methods for expanding speech recognition applications for early resolution of surface safety events (2016)

Summary	**Pros**	**Cons**
• Five potential speech recognition applications were presented, with RCD being the most complex and comprehensive. • Analysis results confirm that context information is valuable for improving speech recognition performance. • Advanced tuning techniques yielded significant speech recognition performance improvement. • Initial findings on the development and performance of a speech recognition system for GC audio are promising. • The FAA must decide on the scope of the vision they want to pursue to plan its investment.	• Speech recognition technology has potential to have a significant positive effect on the detection and prevention of runway incursions. • Context information is valuable for improving speech recognition performance. • Advanced tuning techniques yield significant speech recognition performance improvement. • Initial findings on the development and performance of a speech recognition system for GC audio are promising. • Discussions around the potential value of speech recognition technology were generally positive.	• Speech recognition performance may not currently be sufficient to enable every application. • Development of applications for surface safety can be directly leveraged to develop applications for other purposes and in other domains, requiring the FAA to decide on the scope of the vision they want to pursue.

Table A- 7. Speech recognition for real-time surface safety applications (2018)

Summary	**Pros**	**Cons**
• Demonstrated that the use of a DNN-based speech recognition system can be both more accurate and faster than the previous iteration. • Tested the extensibility of the DNN-based speech recognition system to new facilities, showing that using some facility-specific data is helpful but not necessary to achieve accuracy. • Concluded that ADSB appears technically feasible to be used as a source of surface surveillance that could be paired with speech information to enable intent-based surface safety alerting concepts. • Concluded that existing radar surveillance can be used to predict the surface on which an arrival will land, which can then be compared to speech information to identify and alert controllers to potential wrong surface landings. • Described how speech recognition can enable a smart memory aid device, and recommended the FAA consider it as a technology that could enhance the usefulness of RIDs.	• Demonstrated that the use of a DNN-based speech recognition system can be both more accurate and faster than the previous iteration of the realtime speech recognition platform. • Provided datadriven backing to a requirements development process if the FAA decides to acquire a realtime speech recognition capability. • Broadened the scope of applications that speech can enable • Investigated whether ADSB appears technically feasible to be used as a source of surface surveillance. • Investigated whether existing radar surveillance can be used to predict the surface on which an arrival will land. • Described how speech recognition can enable a smart memory aid device. • Enabled a better understanding of the technology and its use cases. • Progress made toward realtime speech recognition capabilities is directly applicable to using voice data to inform postoperations analysis.	• More research and testing is needed to fully demonstrate feasibility of using existing radar surveillance to predict the surface on which an arrival will land. • No investment decision milestones currently planned for the acquisition of realtime speech recognition capabilities.

Table A- 8. Air traffic control speech recognition (2021)

Summary	**Pros**	**Cons**
• The MITRE Corporation's Center for Advanced Aviation System Development (MITRE CAASD) has been developing and using voice data processing capabilities to analyze ATC radio/voice communications. • These analyses provide new insights to help the Federal Aviation Administration (FAA) make safety and efficiency improvements. • ATC voice communications have a unique set of characteristics that influence technical approaches to voice processing, such as limited frequency bandwidth, fast speech, and domain-specific terminology. • MITRE CAASD has access to FAA voice recordings across 130 FAA ATC facilities, covering most ATC voice transmissions in the NAS. • Enhancements are needed to improve accuracy and add new features to the voice data processing capability to make it useful for new aviation applications and more useful for existing applications. •	• Provides new insights to help the FAA make safety and efficiency improvements. • Uses natural language processing technologies such as automatic speech recognition. • Domain-specific characteristics of ATC speech signal are taken into account. • Voice/speech capabilities are designed to accommodate the unique characteristics of ATC voice communications. • Capabilities are designed to operate on a large scale, handling the majority of voice communications in the NAS • Continues to mature and enhance the processing capabilities to enable more accurate voice analysis.	• Audio of the U/VHF AM radio communication is noisy, with limited frequency bandwidth. • Pronunciation variations, with challenges such as fast speech, coarticulation, and incomplete articulation • ATC vocabulary is limited compared to general English speech. • FAA phraseology is standardized, but controllers and pilots sometimes use local colloquialisms or otherwise vary from standard phraseology. • Identifying speech associated with safety risks can be especially tricky because these risky situations are rare in modern ATC.