Adversarial Attacks on Speech Recognition Systems for Mission Critical Applications: A Survey

Ngoc Dung Huynh[a], Mohamed Reda Bouadjenek[a], Imran Razzak[a], Kevin Lee[a], Chetan Arora[a], Ali Hassani[a], Arkady Zaslavsky[a]

[a]*School of Information Technology, Deakin University, Waurn Ponds Campus, Geelong, VIC 3216, Australia*

Abstract

Machine critical applications are baked into the day-to-day activities of an organization, hence are essential for its survival. The emergence of conversational interfaces in machine-critical applications was long overdue, and we are witnessing how it is slowly but surely becoming commonplace. Recent advances in machine learning, natural language processing, and voice recognition technologies have allowed the development and deployment of speech-based conversational interfaces to interact with various systems and objects such as autonomous vehicles, personal assistants, and various IoT objects. With machine learning growth, recent cyberattacks focus on fooling new visual and conversational Interfaces and have shown their vulnerability to adversarial samples. The fabricated samples trick the deep learning networks to make wrong predictions. This paper investigates the effectiveness of adversarial attacks and defenses against automatic speech recognition in machine-critical applications. Finally, we outlined the challenges and recommendations for adversarial attacks and defense in machine-critical applications.

Keywords: Mission Critical Applications, Adversarial AI, Speech Recognition Systems.

1. Introduction

A mission-critical application, including search and recovery, rescue, military, and emergency management, is a software program, computer, electrical system which is extremely necessary to the success of a business or segment of a business. In other words, a mission-critical application is critical to the survival of a company or organization. Consequently, if a mission-critical application works a little erroneously or interrupted in any way, an enormous negative financial or life-threatening consequence can be immediate. For example, one bank can lose billions if its mission-critical system is defeated, or any disruption in the automatic ambulance locator may impact the rescue operation. Overall speaking, mission-critical applications depend on reliable and timely data. Operators use the information from mission-critical applications to understand the current situation and decide on any needed actions on time.

Email addresses: ndhuynh@deakin.edu.au (Ngoc Dung Huynh), reda.bouadjenek@deakin.edu.au (Mohamed Reda Bouadjenek), imran.razzak@deakin.edu.au (Imran Razzak), kevin.lee@deakin.edu.au (Kevin Lee), chetan.arora@deakin.edu.au (Chetan Arora), reda.bouadjenek@deakin.edu.au (Ali Hassani), arkady.zaslavsky@deakin.edu.au (Arkady Zaslavsky)

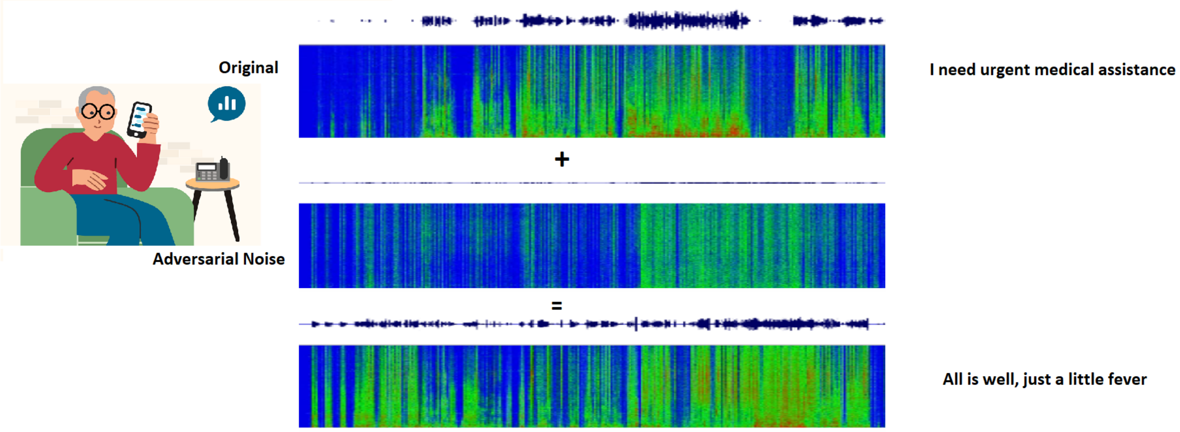
Original
Adversarial Noise
+
=
I need urgent medical assistance
All is well, just a little fever

Figure 1: Illustration of attack on mission critical application: given any waveform, adding a small perturbation makes the result transcribe as any desired target phrase.

Recent technological advancements have revived the concept of direct communication to the devices. Most organizations are swallowed by technological hype in a quest for personalized, efficient, and convenient interaction with customers. A simple but efficient solution is conversational user interfaces. Conversational mission-critical applications are the next great leap forward, with dual-way interaction between machine and human, helping the end user to solve the problems and add the values to the business. Deep learning has been applied to develop the conversational system in mission-critical applications such as safety-critical situations in the banking sector, i.e., deep learning is used to recognize the transcription of medical speech in healthcare [1]. However, the burgeoning success of conversational systems increases the security and privacy concerns with sensitive data. As a result, adversarial attacks in Deep Learning have become one of the most common security threats to mission-critical deep learning applications. Attackers used adversarial examples, inputs to machine learning models to cause the model to make a mistake. For example, attackers fool a face recognition or voice recognition system by adversarial examples to illegally access a financial or government entity.

Voice-enabled conversational systems offer human-like interactions between computers and humans by recognizing our voice and understanding our intentions, deciphering it to other languages, and mimicking human-like conversations. Speech provides a natural way to communicate. Hence, it is the most preferred approach for a conversational system. Speech recognition is the core component of a voice-based conversational system, aiming to convert a speech from an audio form into a textual format for easier downstream processing. In the early days, Hidden Markov Model (HMM) [2] was the primary tool for speech recognition. However, the development of this traditional method has saturated in terms of both latency and accuracy. With Deep Learning (DL) advancements, a neural network replaces some of the standard system components. In recent years, the trend is to design an end-to-end neural network and leverage a massive amount of data to improve the accuracy of speech conversational systems. In the end-to-end models, the modules of the traditional system (acoustic model, pronunciation model, and language model) are jointly optimized in a single system. The examples of end-to-end models are CTC-based models [3] and Attention-based models [4, 5, 6, 7, 8]. Multimodal models [9, 10, 11] have now become more popular since it improves the performance compared to unimodal speech recognition models.

Speech recognition systems have been widely used for mission-critical applications i.e. in the healthcare sector, speech recognition is first used to recognize medical speech between doctors and patients [1]. Besides, speech recognition is also being used to control surgical robots by human surgeons that help to not only save time but also improve the safety of surgeries [12]. Speech recognition systems have also been applied in several other mission-critical domains such as industrial robotics, military applications, and online banking. This led the researcher to develop systems with low latency and high accuracy to accomplish the former tasks, especially in mission-critical infrastructure [13].

Although speech recognition based conversational system (SRCS) has brought many benefits to critical missions, adversarial attacks are the biggest challenge of SRCS, especially Deep Learning-based systems. This problem originated from image classification, and some studies have recently found that adversarial attacks can also target speech recognition components in the conversational system. One of the most popular and effective methods to generated adversarial examples is the Fast Gradient Sign Method (FGSM) introduced by Goodfellow et al. [14]. Originally, FGSM accesses the gradients of a loss function with respect to the input image and then uses the sign of the gradients to generate an adversarial image that obtains the maximized loss. Next, some researchers created adversarial speech examples for speech recognition based on FGSM [15, 16]. After that, Carlini and Wagner [17] used an optimization approach to improve the efficiency of adversarial attacks. Apart from that, the genetic algorithm is also used to fool speech recognition systems [18, 19]. As a result, these attacks significantly reduce the performance of any speech recognition systems. For example, adversarial attacks can lead to severe mistakes in healthcare as shown in Fig. 1. Therefore, it is necessary to improve the efficiency of the speech recognition component to avoid errors that can lead to severe mistakes in mission-critical applications.

This paper aims to review adversarial attacks on the speech recognition component in a conversational system for mission-critical applications. We investigate the effectiveness of adversarial attacks on speech recognition components for mission-critical applications and provide defense techniques against adversarial attacks. Moreover, we also outline the challenges and directions for future research. The paper is organized as follows: Section 2 provides a review of the traditional as well as modern techniques for building speech recognition systems. We then analyze techniques for adversarial attacks in Section 4. In Section 5, we discuss different defense mechanisms against adversarial attacks. Challenges and future directions are discussed in Section 6. Finally, Section 7 concludes the paper.

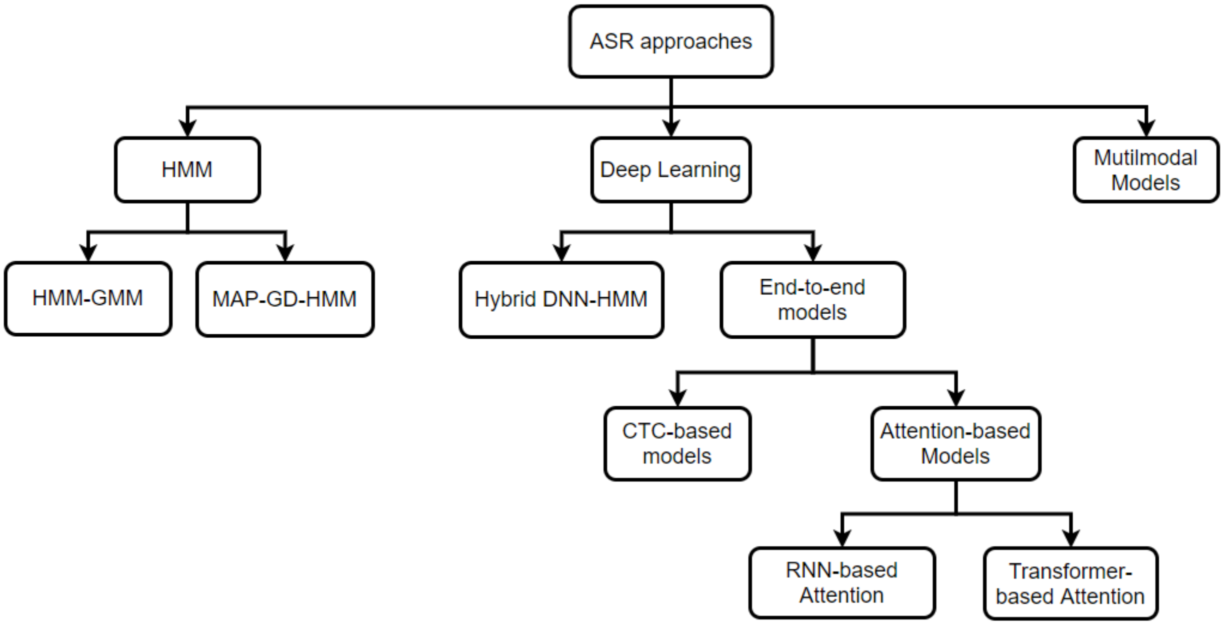

Figure 2: Taxonomy of Speech Recognition.

2. Speech Recognition in Conversational System

Integrating voice recognition technologies with conversational systems opens the door for incredible potential in many real-world applications. Speech recognition provides a natural interface for human communication and is becoming a widely adopted input method for a diverse range of deceives in the smart market. It allows machines to process human voice with the interface and generate the transcription in written form that is sophisticated variation and resembles conservation like an average human. Automatic speech recognition is considered one of the most complex computer science-related mathematics, static, and linguistics domains. The speech recognition system operates in two steps. We first collected and passed to the feature extraction component to extract acoustic features, followed by the decoder to generate the text as the prediction.

2.1. Feature Extraction

The purpose of feature extraction is to convert a speech signal to a predetermined number of frequency components. It is also called Front-end Processing and is implemented by transforming the human speech waveform into parametric representation for subsequent processing and analysis. There are different methods for feature extraction such as Mel Frequency Cepstral Coefficients (MFCC) [20], Perceptual Linear Prediction (PLP) [21], linear predictive coding (LPC) [20], Discrete wavelet transform (DWT)[22], Linear prediction cepstral coefficient (LPCC) [23], Fast Fourier Transform (FFT) [24] and Line spectral frequencies (LSF) [25]. Among these techniques, MFCC is the most popular method for feature extraction [26].

2.2. ***Decoding Approaches***

The aim of decoding is to transcript the audio into words by searching the most likely sequence of words. Searching all possible sequences is astonishingly inefficient. Thus this task is mainly carried out by machine learning such as the Viterbi algorithm (limit the number of searches and finds the optimal path in polynomial time) or deep learning. Earlier, HMM was favored for speech recognition due to its simplicity which is replaced by deep learning due to its flexibility and predicting power. Unlike uni-model deep learning or HMM, recently, multimodal systems (audio plus virtual features) showed better performance. Some popular algorithms of three strategies for the decoder in speech recognition systems are classified in Figure 2.

2.2.1. ***Hidden Markov Model***

Searching all possible sequences is astonishingly inefficient. HMM is a statistical approach to estimating hidden information from visual signals. A speech recognition system is modeled as a Markov process with unknown parameters, signified by the known observable parameters [2]. Integration of Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) outperformed HMM and GMM based system. The decoder in GMM-HMM consists of 3 separately trained modules: acoustic, pronunciation, and language. The acoustic component takes the feature extraction input to predict phonemes using a Gaussian Mixture Model. Pronunciation is an HMM that maps the phonemes expected at the acoustic module to word sequence. The final module is the language module, such as an n-gram language model, which aims to estimates the probabilities of the next word based on the preceding word [2, 27, 28]. GMM-HMM reaches about 21.2% of the word error rate (WER) on the Switchboard data and 35.4% of WER on the CallHome dataset [28]. A complex speech recognition system involves large vocabulary, and word boundaries is challenging to identify. HMM can be less complex and have fewer states for small vocabulary (i.e. one acoustic model per state). However, speech recognition for extensive and continuous vocabulary requires context-dependent modeling, significantly increasing the number of states. Clustering based on a decision tree can be used to shared acoustic models.

GMMs may not be the optimal choice for modeling the distribution of the features of a speech. Another approach to model the distribution is via the maximum of a posterior (MAP) of the generalized Dirichlet (GD)-based HMM (MAP-GD-HMM) [29]. The model MAP-GD-HMM showed better performance compared to HMM-GMM on the TIMIT dataset.

2.2.2. ***Deep Learning***

Deep learning is well known for its applications in image recognition from the last decade, but another developed use of the technology is in speech recognition employed. There are many variations of deep learning architecture for speech recognition because of the flexibility and predicting power of deep neural networks that have recently become more accessible. Some popular architectures based on Deep Learning for speech recognition are Hybrid DNN-HMM and CTC-based models.

Hybrid HMM and Deep Neural Network (NN-HMM) models are the earliest Deep Learning approach where a DNN replaces the acoustic module, and the rest of the modules are kept. The results show that DNN-HMMs can obtain a low WER than HMM-GMMs [30, 31]. However, Long short term memory neural networks (LSTMs) [32, 33, 34] and Time Delay Neural Networks (TDNNs) [35, 36] trend to replaced DNNs and GMMs becoming the acoustic module because these models have shown the improvement of performance in the speech recognition area. In traditional speech recognition systems such as HMM-based models and DNN-HMM, three modules (acoustic, pronunciation, and language models) are separately trained, requiring longer training time and larger memory size. For instance, an n-gram language model requires several gigabytes [27]. In recent years, the trend of building speech recognition systems is to develop an end-to-end speech recognition system based on neural networks. In these models, the modules of the traditional system (acoustic model, pronunciation model, and language model) are trained jointly in a single system. In other words, the network can map the input speech sequence to a sequence (seq2seq) of either graphemes, characters, or words. This approach can resolve the limitation of the traditional system in which the overall system may not be optimal even though individual components are separately optimal. End-to-end architecture systems can be classified into Connectionist temporal classification (CTC)-based models and Attention-based models.

CTC Based models are introduced by [3] is considered as the first architecture of end-to-end models. Speech recognition system based on CTC, also called RNN-CTC models, includes Recurrent Neural Networks (RNNs), which are the main component for sequence processing, and a CTC loss function that arguments the set of target labels with an additional ”blank” symbol. Deep Speech is the popular CTC Based speech recognition [37, 38]. The acoustic and pronunciation modules are trained together in the RNN-CTC model,. However, it still has a limitation: it cannot learn the language module because of conditional independence assumptions. This system uses Markov assumptions to resolve the seq2seq with a forward-backward algorithm [39] followed by Convolutional Neural Networks (CNNs) for RNN-CTC Systems [40, 41].

Generally, CTC-based models still require a language model; attention-based models are built to overcome the limitation of CTC models. Dissimilar CTC-based models, Attention-based models can train traditional speech recognition systems (acoustic, pronunciation, and language modules) directly since they do not have conditional-independence assumptions. Therefore, it does not require a language model that saves memory for the machine [4, 5]. Attention-based models consist of two parts: encoder and decoder (also called attention-based decoder). The purpose of the encoder is to map the acoustic input into a higher representation. Meanwhile, the decoder part takes the output from the encoder part and generates the output symbol based on the full sequence of the preceding predictions. There are 2 major architectures of Attention models: RNN-based encoder-decoders Architecture [4, 5, 6, 7, 8] and Transformer-based encoder Architecture [42, 43, 44].

Recently, RNN based encoder-decoder architecture has been widely used for speech recognition. The encoder, usually a recurrent network, takes the input followed by the decoder, another RNN, only looks at the output of the last hidden state from the encoder. LSTMs are usually as RNNs in the encoder and decoder modules [7]. This architecture has a drawback in that the decoder only sees the last hidden state from the encoder. Therefore, the previous state may not contain the long-range dependencies effectively, leading to the loss of information. To resolve the drawback of the RNN-based architecture, the decoder has access to all the hidden states and relevant input steps [4, 5, 6, 8].

Attention Based RNN is considered one of the best ways to capture the time dependencies in sequences. However, it usually takes a long time to train due to its sequential nature because hidden states are created one step at a time. In addition, the back-propagation for weight updating is also time-consuming. Therefore, a new approach called a Transformer network is designed to resolve the limitation on the RNN-based architecture. This system is still an encoder-decoder architecture, but it does not have ant RNNs or CNNs. Therefore, this system does not require a back-propagation process. Alternatively, the encoder of Transformer Networks consists of 6 identical layers, each of which includes two layers: a multi-head self-attention module, and the second one is a position-wise feed-forward neural network. The decoder also has six identical layers, each containing an additional layer called a multi-head self-attention function. This layer allows Transformer Networks to perform the attention multi-times at the same time. As a result, The training time of the Transformer Network can reduce effectively compared to the RNN-based architecture [42]. The first application of Transformer Network for speech recognition is Speech-Transformer [44]. The author only added some CNNs before putting features to the input to the Transformer Network to reduce the difference in the dimensions of the input and output sequences. As a result, it reached about 10.9% of WER on the Wall Street Journal (WSJ) speech recognition dataset. After that, this architecture is improved by integrating CTC loss into Speech-Transformer. As a result, the WER reduces to 4.5% on WSJ. the authors point out that the Transformer Network obtains higher accuracy and shorter training time than the RNN-based Attention model [43].

2.3. ***Mutilmodal models***

The integration of multiple modalities while talking helps people understand each other better. In other words, combining the context with our conversation helps to convey more accurately what we mean. Generally, a unimodal speech recognition system is trained on the data that contains a speech as the input and a sequence of words as the labels. However, the input to a multimodal model includes both speech and image. There are two main steps in this approach. The first step is to extract the audio and visual features. The second step is to integrate these features into one vector to train the model [9]. For example, Mamyrbayev et al. [45] combine human voices and images of the lip, face, and gestures for speech recognition. As a result, multimodal models can improve the accuracy compared to unimodal models. Therefore, multimodal speech recognition models that combine audio and visual modalities have become more prevalent in speech and natural language processing communities.

Like general speech recognition, the are several different methods to extract audio features, such as MFCC. For image feature extraction, the general method is using CNNs such as ResNet [9, 10, 11], and Region Convolutional Neural Network (RCNN) [46]. The most important of this system is the method to combine the audio features and images features. This process can be implemented in both encoder and decoder. For encoder feature fusion, we can apply the technique called Shift Adaptation [11]. For decoder feature fusion, the most popular method uses Early Decoder Fusion to integrate these features [9, 11, 46]. Apart from that, Weighted Early Decoder Fusion, Middle Decoder Fusion, and Hierarchical Attention over Features can also be used to combine the audio and image features [11]. The result showed that multimodal systems can obtain lower WER compared to unimodal models, especially under noise conditions [9, 10, 11, 46, 45]. Additionally, Weighted Early Decoder Fusion achieves the lowest WER, increasing 1.40% on the augmented dataset. Finally, hierarchical Attention over Feature achieves the best recovery rate of masked words, with an improvement of 4% over an attention-based model [11].

3. Datasets and Evaluation Metrics

3.1. Speech Dataset

A dataset is essential for the training and testing process of a speech recognition system. Several different commonly used source is presented as follow:

- *Medical dataset*: Unlike every day, speech medical dictations are usually slow and restarted sentences [1]. Besides, speech medical dictations consist of domain-specific medical terminology, including thousands of drug names. Medical dataset [1] consists of 270 hours of medical speech data that doctors and patients generate.

Apart from health sector data, data on specific areas such as banking and military are not available or recorded by their staff. They are still used normal speech to create speech recognition systems. Some popular normal dataset is listed as follow:

- *Switchboard*: The Switchboard-1 Telephone Speech Corpus is an open used source for speak recognition [49]. The dataset consisted of approximately 260 hours of speech and was collected by 543 speakers (302 male, 241 female) by Texas Instruments in 1990. The dataset covers Seventy topics covered in the dataset, of which about 50 were used frequently.

- *CallHome*: The CallHome dataset is an American English speech data that consists of 18.3 hours of transcribed spontaneous speech, comprising about 230,000 words.

- *TIMIT*: The TIMIT acoustic-phonetic continuous Corpus is a common dataset that includes broadband recordings of 6300 phonetically rich sentences [50]. 30% of the speaker are female, and the rest are male speakers. The training set consists of 3.14 hours of recording.

Table 1: Comparison of speech recognition systems.

speech recognition system	**Dataset**	**WER (%)**	**Accuracy(%)**
HMM-GMM [28]	Switchboard	21.2	
HMM-GMM [28]	CallHome	36.4	
HMM-GMM [29]	TIMIT		50
MAP-GD-HMM [29]	TIMIT		93.33
DNN-HMM [28]	Switchboard	14.2	
DNN-HMM [28]	CallHome	25.7	
LSTM-HMM [47]	Switchboard	7.2	
LSTM-HMM [47]	CallHome	12.7	
TDNN-HMM [35]	Switchboard	9.2	
TDNN-HMM [35]	CallHome	17.3	
LSTM-CTC (Bigram Language Model) [39]	Wall Street Journal	13.5	
LSTM-CTC (No linguistic information) [39]	Wall Street Journal	27.3	
LSTM-CTC (Trigram language model) [39]	Wall Street Journal	8.2	
LSTM-CTC[39]	Wall Street Journal	8.2	
CNN-LSTM-TCT [40]	Wall Street Journal	10.5	
LSTM-CTC [48]	Medical Dataset	20.1	
Attention RNN-based [4]	Wall Street Journal	8.0	
Attention RNN-based [1]	Medical Dataset	15.4	
Attention RNN-based [48]	Medical Dataset	18.3	
Transfomer Network [44]	Wall Street Journal	10.9	
Transfomer Network [43]	Wall Street Journal	4.5	
Unimodal Model (Attention RNN-based) [11]	Flickr8K	13.7	
Multimodal [11]	Flickr8K	13.4	

- *Wall Street Journal*: The Wall Street Journal (WSJ) is the English speech source with an extensive vocabulary, natural language, high perplexity [51]. The dataset contains 400 hours of speech data and 47,00 text data. This dataset is usually used for speech recognition and natural language processing.
- *Flickr 8K*: The Flickr 8k Audio Caption Corpus dataset 40,000 spoken captions of 8,000 natural images that capture the actions of people or animals. This dataset is commonly used to build multimodal speech recognition systems.
- *Common Voice*: It is the open-source voice regularly updated dataset created by Mozilla [52]. Common Voice dataset consists of 13,905 recorded hours of speech. The categories of the dataset are divided into demographic metadata like age, sex, and accent. The dataset contains 11,192 validated hours in 76 languages such as English, India, and South Asia.
- *LibriSpeech speech recognition corpus*: LibriSpeech created Panayotov et al. [53] is a corpus of approximately 1000 hours of 16kHz read English speech.

3.2. Evaluation Metric-done

The evaluation of attack is subjective as input perturbation varies; thus human eye can not distinguish while still impact the classification performance. An adversarial attack can be harmful and worthy to investigate if it is effective and efficient. An attack can be considered effective if the attack can not be differentiated from the original action and negatively impact the performance (accuracy). An attack can be considered efficient if it requires minimal resources. From the perspective of the speech recognition system, effectiveness is mostly considered criteria by determining the accuracy score of the speech recognition system. As the output of systems may not have the same length as the target thus, it is hard to compute the accuracy. ***Word Error Rate*** is the more popular used metric to evaluate speech recognition systems. The WER can be computed as the following formula:

$$WER = \frac{S + D + I}{N} \tag{1}$$

Where S is the number of substitutions performed in the prediction compared to the reference; D is the number of deletions; I is the number of insertions; N is the number of words in the reference. As WER calculates the error of the speech recognition system, the low WER indicates the better speech recognition systems.

4. Adversarial Attacks Techniques-working

Speech recognition systems are becoming popular in mission-critical applications such as healthcare [12], bank, and military [13]. Conversational systems such as chatbots or virtual agents are in demand with the progress in speech recognition for mission-critical applications. However, despite its benefits, speech recognition systems are vulnerable to potential security risks, especially adversarial attacks.

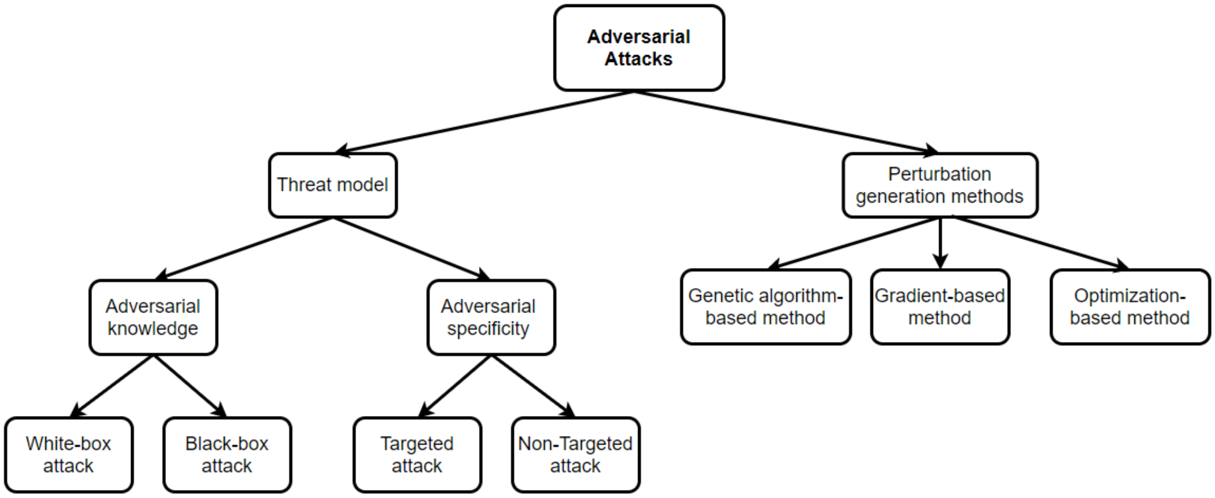

Figure 3: Taxonomy of Adversarial Attacks.

Adversarial attacks consist of subtly modified original audio in a way that slight change is unnoticeable to the human ear, however, can impact the machine learning model significantly and results in very poor recognition. Adversarial attacks aim to generate adversarial samples to reduce the performance of speech recognition systems. However, preceding research focuses on the development of adversarial attacks and their defense for image classification tasks; however, speech recognition systems are considerably less focused. Deep learning-based speech recognition has shown vulnerability against adversarial examples. Adversarial attacks create non-random imperceptible perturbations as adversarial examples and add them in input examples via optimization algorithms. As a result, the speech recognition system is fooled and makes wrong decisions [12]. Figure 1 illustrates the impact of adversarial attacks in critical applications. In the following discussion, we reviewed the adversarial attacks on speech recognition from two aspects: the threat model and the methods to generate adversarial examples (Figure 3).

4.1. Threat Model

Adversarial attacks can occur at the deployment/testing stage of the victim model, and they can not be used directly for attacking because of entirely different threat models. According to the background, knowledge, and objective, threat adversary models can be divided into Adversarial Knowledge and Adversarial Specificity.

Adversarial Knowledge It defines the level of available information to the adversary. *White-box attacks:* Attackers have full knowledge of the network, can directly access the target network, and use the models' information such as the model parameters, model architecture, model type, and training weights to generate perturbation through loss minimization over attacking object, target object, and perturbation. The attack can be considered if the network's output is the target object regardless of how perturbations are added in the audio.

Table 2: Comparison of adversarial attack methods

Author	**Attack Model**	**Target Model**	**Adversarial Knowledge**	**Adversarial Spesificity**	**Note**
Gong & Poellabauer [15]	Gradient sign	Wave CNN	White box	Non-Targeted	WER increases from 12% to 25%
Kreuk et al. [16]	Gradient sign	Customized RNN	White box	Targeted	The accuracy reduced from 81% to 62.25%
Alzantot et al. [18]	Genetic algorithm	Customized CNN	Black box	Targeted	87% targeted attack success rate and the noise does not change 89% the human listener's perception
Taori et al. [19]	Genetic algorithm	DeepSpeech	Black box	Targeted	35% targeted attack success rate
Cisse et al. [54]	Optimization	DeepSpeech2	White box and Black box	Non-Targeted	Reduce 12% of WER
Carlini and Wagner [55]	Optimization	DeepSpeech	White box	Targeted	100% attack success rate and 99% similarity between adversarial speech and original speech
Wang et al. [56]	Optimization	DeepSpeech2	White box	Non-Targeted	98% attack success rate
Quin et al. [57]	Optimization	Lingvo	White box	Targeted	100% targeted attack success rate on arbitrary full-sentence targets
Zelasko et al. [58]	Gradient sign and Optimization	DeepSpeech2 Transformer	White box	Targeted	Under FGSM, WER of DeepSpeech2 and Transformer increases by 78% and 7%, respectively. Under Imperceptible attack, WER of two systems increases by $4-5\%$. Only DeepSpeech2 is broken by PGD

Black-box Attack: Unlike white-box attacks, the model might be accessible to the attacker. In this case, the attacker acts as a regular user and queries the model with input audio to get the result. Attacker works only on perturbation and uses gradients of a misclassification objective to achieve target output.

Untergeted Attack It may possible that attacker wants only the output to be incorrectly classified and does not have particular targeted output. To generate such attacks, attacker considers to maximizes the distance between label and perturbed output.

Adversarial Specificity It establishes what is the aim of an adversarial attack. *Non-target attacks:* Attacks perturbs the audio in such a way that it does not assign a specific class to the output. The prediction can be randomly output except for the actual class. As a result, this type of attack has more options to mislead the output.

Targeted attacks: Attacks mislead the acoustic system to predict a specific target that the adversary decides. This kind of adversarial attack is more powerful because it is more realistic and challenging than non-target attacks.

4.2. Perturbation generation method

Based on the techniques to generate perturbations in speech recognition systems, the adversarial methods are divided into three options: (i) genetic algorithm-based method, (ii) gradient sign-based method, and (iii) optimization-based method.ods are divided into three options: (i) genetic algorithm-based method, (ii) gradient sign-based method, and (iii) optimization-based method

4.2.1. Genetic algorithm-based method

A genetic algorithm is a model inspired by Charles Darwin's theory of natural evolution that requires three steps: population, selection, and mutation. Firstly, a random population that contains adversarial examples is created, followed by a fitness function to assign the adversarial examples with higher fitness to mutate and become part of the next generation. Finally, the sequence of actions will be repeated over and over until the desired result is achieved.

The first adversarial attack based on a Genetic algorithm on speech recognition was proposed by Alzantot et al. [18]. By following the primary step of the genetic algorithm, they have created a targeted black-box attack with 87% of success rate and 89% of participants still recognize the adversarial audio as the original audio. However, this study has few limitations, such as the speech recognition system can only recognize single words, and the attacks were only forceful on a customized CNN. As a result, the proposed adversarial attacks may not effectively work for new speech recognition systems.

Taori et al. [19] combined a genetic algorithm and a gradient estimation to improve the existing attack. First, they used a genetic algorithm to find suitable examples on the population of candidates followed by gradient estimation to explore more noise when the adversarial examples are nearing their target. Although the convergence process is faster than the original method [18], however, the efficiency of the attacks is poor with the success rate of the attack 35%. Besides, the similarity rate, which candidates the distance between the output and the target, achieved 89.25%.

4.2.2. Gradient sign based method

In this method, the attackers will access the information of the model to generate adversarial examples by using the fast gradient method (FGSM). More specifically, FGSM uses the gradient of the targeted model to find adversarial examples. Then, the speech input is manipulated by adding small noises to become the adversarial examples. The advantage gradient sign is that it can faster generate adversaries because it bases on one round of iteration to adjust speech features.

Gong and Poellabauer [15] introduced the first adversarial attack using the gradient sign-based method. The authors proposed an end-to-end approach by directly adding perturbations to the raw waveform. The problem of the Gradient sign-based methods is the vanishing problem, in which the gradients of the loss function approaches zero, making the network hard to train. To resolve this issue, They replaced the recurrent layers with convolutional layers. The result shows that the WER of the speech recognition system increases from 29% to 44% on WaveCNN. However, assuming that the adversarial attack knows the model is unrealistic in practice. Additionally, the transferability of adversarial examples was not explored. Similarly, Kreuk et al. [16] applied the same method over acoustic features and then reconstructed the audio waveform from adversarial acoustic features. To explore the transferability of the adversarial example, they create two black-box attacks on two different models. The authors pointed out that their attack is efficient because it reduces the accuracy of the speech recognition system from 81% to 62.25% and increases the false positive rate from 16% to 46%. However, the same problem with the former work [15] is assuming that the adversarial attack knows the model is unrealistic in practice. Apart from that, the magnitude of perturbation was not explored. Only ABX experiments were conducted to assess detectable differences of adversarial spectrogram examples.

4.2.3. Optimization based method

Optimization methods require complete knowledge of the model, such as model architecture and model parameters. Unlike the gradient-based methods, the optimization-based method iteratively uses FGSM to find a minimum perturbation for the target input. DeepFool [59] and Projected Gradient Descend (PGD) [60] are the popular attacks based on the Optimization-based method. Such method can create smaller perturbations than in FGSM. As a result, this approach is more powerful than other approaches since it usually reaches a higher attack success rate.

Optimization-based CW attacks are most powerful as they are more imperceptible and have less distortion in the produced attack than other methods. Besides, the implementation of CW attacks can sometimes be tricky and needs to select parameters efficiently to obtain the desired data. Cisse et al. showed a flexible adversarial attack named Houdini [54] based on DeepFool's structure. The attack gets the loss value between true target and predictions and then the forward-backward process to seek the adversarial examples on DeepSpeech2. The result shows that the adversarial attack causes misclassification with 12% of WER on the Librispeech dataset; however, the exact perturbation was not investigated.

Carlini and Wagner [55] presented an adversarial target attack that directly adjusts raw waveform via an optimization approach. They demonstrated a novel approach to change via the entire network that can achieve faster convergence and lower perturbation magnitude. The result showed that the adversarial speech is 99% like the original speech, and the attack success rate is 100% on DeepSpeech. Nevertheless, assuming that the attacker knows the model is unrealistic in practice. Besides, the transferability of adversarial examples and the proposed technique's applicability over the air are not explored. Quin et al. [57] improved the construction of adversarial examples generated by Carlini and Wagner [55] by using the psychoacoustic principle of auditory masking. They have used the optimization method with two stages: the first stage explores a perturbation to fool the target network, and the second stage optimizes the perturbation to make it imperceptible to humans. As a result, the model can generate more imperceptible and robust adversarial examples and achieve a 100% attack success rate on Lingvo, a state-of-the-art sequence-to-sequence speech recognition model. Nevertheless, assuming adversarial examples can access the model is unrealistic.

Recently, Wang et al. [56] proposed a novel and effective attack on speech recognition systems named (SGEA), which includes four stages: mini-batch gradient estimation, iterative momentum method, coordinate selection, and batch size adaptation. The mini-batch gradient estimation separates the number of queries required in each iteration from the high input dimensions, which creates a balance between the convergence speed and the number of queries per iteration. After that, the iterative momentum method adds perturbations to the approximated gradient followed by coordinate selection to increase the converging speed. Finally, the last stage automatically and effectively acquires the appropriate batch size value for each audio. Results showed that SGEA achieved 98% attack rate.

Recently, Zelasko et al. [58] used FGSM, PGD, and the imperceptible attack proposed by Quin et al. [57] to attack the two most effective speech recognitions (DeepSpeech2 and Transformer encoder-decoders speech recognition). They point out that both speech recognitions systems are also vulnerable to three adversarial attacks. Under FGSM, the WER of Deepspeech2 increases by 78%, while the WER of Transformer increases by only 7%. Under PGD, DeepSpeech2 is completely broken, while it is not practical to fool Transformer. Additionally, both systems can be attacked by the Imperceptible attack since the WER rises by from 4 - 5%. Therefore, Transformer is less affected by these attacks than DeepSpeech2. Finally, they claim that the transferability of adversarial attacks is limited.

5. Defenses against adversarial attacks

Adversarial examples are a massive threat to mission-critical applications and can result major disaster. Thus, defense against adversarial attacks for machine critical application is important than any other application due to sensitivity of the problem. There are three key goals of adversarial defenses:

- Reduce the impact of adversarial examples on the model architecture.
- Maintain the training time and accuracy of the model.
- Focus on the adversarial examples that are close to the original training examples.The reason is that the examples which are far from the training examples are secure enough.

Based on the aims of defenses against adversarial attacks, defense techniques for speech recognition systems can be divided into two categories: *Reactive Defenses* and *Proactive Defenses*.

5.1. Reactive defense

Reactive defenses approach relies entirely on being to able to shore up the defense before an attacker attempt and exploit vulnerability or may alarm if security is breached by detecting adversarial examples and normal examples. It keeps the organization in continuous firefighting mode. Reactive defenses is classified into 2 techniques:*Adversarial Detecting* and *Network Verification*.

5.1.1. Adversarial Detecting

The adversarial detecting method can be considered as a binary classification to classify adversarial examples and normal samples. The input of the classifier can be either acoustic features or raw waveform. Generally, binary classification based detection may achieve higher accuracy for detecting adversarial examples. However, it requires classifier for each threat, which costs time and memory to train the classifier.

Rajaratnam et al. [61] presented a classifier to detect the adversarial examples introduced by Alzantot et al. [18]. The author applied processing models such as Band-pass Filtering, Speech compression, and AAC compression to detect the adversarial examples. Beside, they also use several ensemble detection models such as Majority Voting, Random Forest Classification, and Extreme Gradient Boosting. The result showed that their best model's precision and recall scores are 93.5% and 91.2%, respectively. Samizade et al. [62] performed a CNN classification on spectral features on the Google Speech Command dataset. They aimed to detect the adversarial examples generated by Alzantot et al. [18]. Their model was very successful since the detection accuracy reaches approximately 100%.

5.1.2. Network Verification

Recently, researchers have focused on verification method for detection of adversarial attacks through counterexample and advances have already been made. Network verification finds adversarial examples based on the different samples. For instance, we use an speech recognition system to generate various transcriptions. After that, we compute the flood scores of adversarial examples and benign examples. Finally, we use a specific threshold to detect adversarial examples if the flood scores are less than the threshold.

Rajaratnam and Kalita [63] presented a novel adversarial detection by adding random noise to audios. Consequently, they calculate the flood scores of adversarial examples and original speech. Finally, examples with less than a specific score of flood scores will be marked as adversarial examples. The experimented result showed that the detection precision and recall score are high with 91.8% and 93.5%, respectively. Ma et al. [64] used the same technique for adversarial detection on a multimodal speech recognition model. First, they use the temporal correlation between the audio feature and video features. The second step is to compare the correlation value with a specific threshold. If the correlation value is less than the threshold, they assign the speech as an adversarial example. Their result shows that their model's detection precision, recall, and f1 score are 91.8%, 93.5%, and 92.6%, respectively.

5.2. Proactive Defense

Proactivity methods involves identification and mitigation of hazardous conditions to enhance the robustness when building speech recognition systems. Based on different techniques, the proactive defenses are divided into two categories: *Adversarial Training* and *Robustifying Model*:

5.2.1. Adversarial Training

Unlike most of the adversarial attack detection models are found to be less effective, adversarial training are efficient for defense against attacks. Adversarial Training resists adversarial examples by retraining the model with adversarial examples. Even though, adversarial training is popular defense approach against the existing adversarial examples. however may not be effective with unknown attacks. Sun et al. [65] integrated the adversarial example created by FGSM to the training set to retrain the speech recognition model. Additionally, they also used an algorithm called Teacher/Student training to make the model more robust. Their proposed adversarial training reduces WER by 23% on the Aurora-4 single task.

5.2.2. Robustifying Models

Instead of considering to achieve high accuracy, excluding non-robust features and robustifying the latent space may guide the model to learn robust features which can be achieved by dual manifold adversarial training i.e. adding crafted adversarial examples to the audio training set as well as latent space o make the model robust against similar attacks.

Esmaeilpour et al. [66] combined the pre-processed discrete wavelet transform representation of audio signals and Support Vector Machine to secure audio systems against adversarial attacks. The author used a neural network to smooth the spectrograms to reduce the impact of adversarial examples. The smoothed spectrograms were processed by dynamic zoning and grid shifting using the speeded-up robust features (SURF), which transform into cluster distance distribution using the K-Means++ algorithm. The output is then fed into an SVM. The result shows that the proposed method can provide a good trade-off between accuracy and resilience of the most adversarial examples generated by BackDoor and Dolphine Attack.

Tamura et al.[67] presented a novel approach based on a sandbox approach to eliminate adversarial examples. At first, the objective is eliminate the perturbation in adversarial examples by eliminating techniques such as dynamic down-sampling and denoising. After that, they compare the characteristic error rate of transcription results of DeepSpeech and then regard the adversarial examples that obtain a more significant characteristic error rate than the threshold.

Zelasko et al. [58] use three different defenses against FGSM, PGD, and the imperceptible attack: Randomized smoothing, WaveGAN vocoder, and Label smoothing. Firstly, Randomized smoothing aims to map adversarial signals with additive random, normally distributed noise that controls the trade-off between robustness and accuracy. Randomized smoothing is considered as a defense method that against norm-bounded adversarial examples. Meanwhile, they WaveGAN [68] as a processing defense to reconstructs the log-Mel-spectrograms. This method can enhance the stability and efficiency of adversarial training. Finally, Label smoothing is a regularization technique that introduces noise for the labels when the loss function is cross-entropy. The model applies the soft-max function to the penultimate layer's logit vectors to compute its output probabilities. The result shows that, among three methods, Randomized smoothing is the most effective technique against adversarial attacks on DeepSpeech2 and Transformer. Other the other hand, WaveGAN vocoder can reduce the success attack rate of adversarial examples. However, adversarial attacks can access the WaveGAN structure and fool WaveGAN and speech recognition at the same time. Finally, without label smoothing, speech recognition could be more vulnerable.

6. Challenges and Future Directions

Attackers are considering gradient or nuance from defensive techniques to generate perturbations. In machine critical applications, we seek a secure speech recognition. A model that provides efficient performance most of the time can be safe for other speech recognition task however, is not safe for machine critical applications. Even certified defense system can be broken if heavy disturbing is applied. Hence, no defense method can be claimed to be effective for new threats. Adversarial attacks [56, 57, 55, 19] are the earliest effective attack, however, recent threats have opened up new opportunities that need to be resolved in the future. There are following three main challenges:

Transferability: The transferability of adversarial examples is considered an effective approach against adversarial attack i.e., the adversarial examples created for Model A can flood Model A and can be effective to attack model B. In contrast, model A and model B do not have the same structure. The transferability of adversarial examples has been mainly exploited in the image domain. However, the transferability of speech adversarial examples is not yet widespread. For example, the Houdini method proposed by Cisse et al. [54] only presents that the adversarial examples generated by a DeepSpeech2 system can effectively attack the Google voice system. Additionally, the method proposed by Kreuk et al. [16] illustrates that adversarial examples can keep up the transferability between two models that are trained on two different datasets with the same architecture. Therefore, it is considered that adversarial examples can achieve transferability in a specific set of models. However, whether adversarial examples can attack arbitrary target models needs to be explored in the future. One solution is to research deeper into the theory of famous adversarial examples to find generalization perturbation.

Played Over-the-Air: The adversarial examples for speech recognition systems are considered a real adversarial attack if the targeted recognition is produced when the signal is played over the air. However, in preceding research, the speech signal is directly fed into the system, which is unrealistic in over-the-air attacks. Cisse et al. pointed out the potency of over-the-air adversarial examples. However, because they only proposed non-targeted adversarial attacks, they can easily succeed. Additionally, the proposed method requires the background sound to be quiet that is unrealistic because the environment around us is usually noisy. Quin et al.[57] used an acoustic room simulator to make progress towards physical-world over-the-air audio adversarial examples. Nevertheless, this approach is not full over-the-air. Therefore, creating played-the-air speech adversarial examples in a real-world environment is an open research topic for future research.

Target Multiple Inputs: Almost all adversarial examples aim to attack waveforms, so very few attacks focus on other input features such as spectrogram and MFCC features. Adversarial attacks should focus on different target features to determine whether adversarial attacks on speech recognition flood more effective on wave-from than on other target features. Additionally, the effectiveness of attacks on multiple input types also needs to be further investigated. Apart from some challenges and future directions for adversarial examples, we point out some difficulties and recommendations of adversarial defenses as following:

Reactive Defenses: Adversarial detecting is an effective method to detect adversarial examples. The core of adversarial detecting is classifying the input transformation into two categories: adversarial and normal examples. For future research, ensemble methods can be used to detect adversarial examples more effectively. On the other hand, Network Verification is the new technique that shows effectiveness in detecting attacks. In the future, this domain should be focused on exploring the understanding and characteristics of adversarial examples.

Proactive Defenses: It aims to create a robust network to prevent adversarial examples. Adversarial Training and Robustifying Model are promising methods to resist the adversarial examples. Inspired by the image domain, Adversarial Training is a promising approach used for the robustness of the conversational system. On the other hand, using denoising techniques such as GAN [69] to eliminate the adversarial perturbation is also an encouraging method to prevent the adversarial examples. Proactive defense is invalid against white-box attacks or grey-box attacks, whereas reactive defense is sensitive to the transferability of adversarial examples or low distortion adversarial examples.

7. Conclusion

With the progress in machine learning, conversational systems have been actively deployed in real-world applications. With the emergence of conversational system, cyber attacker are active in development of methods to foll the new visual and conversational Interfaces and have shown their vulnerability to adversarial samples. In this paper, we analyzed popular algorithms used to create speech recognition systems, such as Hidden Markov Model and Neural Network. We also provide a comprehensive view of adversarial examples and the method used to defense these negative examples. Finally, through the limitations of adversarial examples, we introduce the potential challenges in speech adversarial examples and provide future researches for these challenges.

References

[1] Erik Edwards, Wael Salloum, Greg Finley, James Fone, Greg Cardiff, Mark Miller, and David Suendermann-Oeft. Medical speech recognition: Reaching parity with humans. *Springer International Publishing*, pages 512–524, 2017.

[2] Monica Franzese and Antonella Iuliano. An introduction to hidden markov models. *IEEE ASSP Magazine*, 3(1):4 – 16, 1986.

[3] Alex Graves, Santiago Fernandez, and Khudanpur Sanjeev. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

[4] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[5] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. End-to-end attention-based large vocabulary speech recognition. *arXiv:1508.04395*, 2016.

[6] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. *arXiv:1506.07503v1*, 2015.

[7] Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. End-to-end continuous speech recognition using attention-based recurrent nn: First results. *arXiv:1412.1602*, 2014.

[8] Kyunghyun Cho, Bart Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv:1406.1078*, 2014.

[9] Tejas Srinivasan, Ramon Sanabria, and Florian Metze. Looking enhances listening: Recovering missing speech using images. *arXiv:2002.05639*, 2020.

[10] Ozan Caglayan, Ramon Sanabriay, Shruti Palaskary, Loic Barrault, and Florian Metze. Multimodal grounding for sequence-to-sequence speech recognition. *arXiv:1811.03865v2*, 2019.

[11] Tejas Srinivasan, Ramon Sanabria, Florian Metze, and Desmond Elliott. Multimodal speech recognition with unstructured audio masking. *arXiv:1511.03690*, 2020.

[12] Siddique Latif, Junaid Qadir, Adnan Qayyum, Muhammad Usama, and Shahzad Younis. Speech technology for healthcare: Opportunities, challenges, and state of the art. *IEEE Reviews in Biomedical Engineering*, 14:342–356, 2021.

[13] Jayashri Vajpai and Avnish Bora. Industrial applications of automatic speech recognition systemsint. journal of engineering research and applications. *Int. Journal of Engineering Research and Applications*, 2016.

[14] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv:1412.6572*, 2015.

[15] Yuan Gong and Christian Poellabauer. Crafting adversarial examples for speech paralinguistics applications. *arXiv:1711.03280*, 2017.

[16] Felix Kreuk, Yossi Adi, Moustapha Cisse, and Joseph Keshet. Fooling end-to-end speaker verification by adversarial examples. *arXiv:1801.03339*, 2017.

[17] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. *arXiv:1608.04644*, 2016.

[18] Moustafa Alzantot, Bharathan Balaji, and Mani Srivastava. Did you hear that? adversarial examples against automatic speech recognition. *arXiv preprint arXiv:1801.00554*, 2018.

[19] Rohan Taori, Amog Kamsetty, Brenton Chu, and Nikita Vemuri. Targeted adversarial examples for black box audio systems. *arXiv:1805.07820*, 2019.

[20] Pavel Paramonov. Fast algorithm for isolated words recognition based on hidden markov model stationary distribution. *2017 IEEE 4th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, 2017.

[21] Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *The Journal of the Acoustical Society of America*, 87(4), 1990.

[22] Akshay Chamoli, Ashish Semwal, and Nomita Saikia. Detection of emotion in analysis of speech using linear predictive coding techniques (l.p.c). *International Conference on Inventive Systems and Control (ICISC)*, 2017.

[23] Aadel Alatwi, Stephen So, and Kuldip K. Paliwal. Perceptually motivated linear prediction cepstral features for network speech recognition. *10th International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2017.

[24] Weiqiang Liu, Qicong Liao, Fei Qiao, Weijie Xia, Chenghua Wang, and Fabrizio Lombardi. Approximate designs for fast fourier transform (fft) with application to speech recognition. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 66(12):4727 – 4739, 2019.

[25] Himadri Mukherjee, SKMD Obaidullah, Obaidullah Santosh, Santanu Phadikar, and Kaushik Roy. Line spectral frequency-based features and extreme learning machine for voice activity detection from audio signal. *International Journal of Speech Technology*, 21:753–760, 2017.

[26] Sabur Ajibola Alim and Nahrul Khair Alang Rashid. *Some Commonly Used Speech Feature Extraction Algorithms*. Intech Open, 2018.

[27] Daniel Jurafsky and James Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (Third Edition draft)*. Pearson, 2020.

[28] Karel Veselỳ, Arnab Ghoshal, Lukás Burget, and Daniel Povey. Sequence-discriminative training of deep neural networks. In *Interspeech*, volume 2013, pages 2345–2349, 2013.

[29] Samr Ali and Nizar Bouguila. Maximum a posteriori approximation of hidden markov models for proportional sequential data modeling with simultaneous feature selection. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12, 2021.

[30] Andrew Maas, Peng Qik, Ziang Xie, Awni Hannun, Christopher Lengerich, Daniel Jurafsky, and Andrew Ng. Building dnn acoustic models for large vocabulary speech recognition. *arXiv:1406.7806*, 2015.

[31] Mohamed A, G G, Hinton, and Penn G. Understanding how deep belief networks perform acoustic modelling. *IEEE International Conference on Acoustics*, 66(12):4273–4276, 2012.

[32] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. Hybrid speech recognition with deep bidirectional lstm. *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013.

[33] Larkin Liu, Yu-Chung Lin, and Joshua Reid. Improving the performance of the lstm and hmm model via hybridization. *arXiv:1907.04670v4*, 2021.

[34] Wei Zhou, Ralf SchlÃŒter, and Hermann Ney. Full-sum decoding for hybrid hmm based speech recognition using lstm language model. *arXiv:2004.00967*, 2020.

[35] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Interspeech*, pages 2751–2755, 2016.

[36] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth annual conference of the international speech communication association*, 2015.

[37] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Ng. Deep speech: Scaling up end-to-end speech recognition. *arXiv:1412.5567*, 2014.

[38] Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher

Fougner, Tony Han, Awni Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, and Zhenyao Zhu. Deep speech 2: End-to-end speech recognition in english and mandarin. *arXiv:1512.02595*, 2015.

[39] Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772. PMLR, 2014.

[40] Y Zhang, W Chan, and N Jaitly. Very deep convolutional networks for end-to-end speech recognition. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4845–4849, 2017.

[41] Brendan Shillingford, Yannis Assael, Matthew Hoffman, Thomas Paine, CÃan Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, Marie Mulville, Ben Coppin, Ben Laurie, Andrew Senior, and Nando Freitas. Large-scale visual speech recognition. *arXiv:1807.05162*, 2018.

[42] A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A Gomez, L Kaiser, and I Polosukhin. Attention is all you need. *arXiv:1706.03762v5*, 2017.

[43] Shigeki Karita, Nelson Soplin, Shinji Watanabe, Marc Delcroix, Atsunori Ogawa, and Tomohiro Nakatani. Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration. *INTERSPEECH 2019*, 2019.

[44] Linhao Dong, Shuang Xu Xu, and Bo Xu. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[45] Orken Zh Mamyrbayev, Keylan Alimhan, Beibut Amirgaliyev, Bagashar Zhumazhanov, Dinara Mussayeva, and Farida Gusmanova. Multimodal systems for speech recognition. *International Journal of Mobile Communications*, 18(3):314–326, 2020.

[46] David Harwath and James Glass. Deep multimodal semantic embeddings for speech and images. *arXiv:1511.03690*, 2015.

[47] George Saon, Gakuto Kurata, Tom Sercum, Samuel Audhkhasi, Kartik amd Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Michael Ramabhadran, Bhuvana amd Picheny, Lynn-Li Lim, Bergul Roomi, and Phil Hall. English conversational telephone speech recognition by humans and machines. *arXiv:1703.02136*, 2017.

[48] Chung-Cheng Chiu, Anshuman Tripathi, Katherine Chou, Chris Co, Navdeep Jaitly, Diana Jaunzeikare, Anjuli Kannan, Patrick Nguyen, Hasim Sak, Ananth Sankar, Justin Tansuwan, Nathan Wan, Yonghui Wu, and Xuedong Zhang. Speech recognition for medical conversations. *arXiv:1711.07274*, 2018.

[49] John Godfrey and Hollimanr Edward. Switchboard-1 release 2. 1993.

[50] JS Garofolo, LF Lamel, WM Fisher, JG Fiscus, DS Pallett, NL Dahlgren, and Victor Zue. Timit acoustic-phonetic continuous speech corpus. 1993.

[51] JS Garofolo, David Graff, Doug Paul, and David Pallett. Csr-i (wsj0) complete. 1993.

[52] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

[53] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[54] Moustapha Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured prediction models. *http://arxiv.org/abs/1707.05373*, 2017.

[55] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. *arXiv:1801.01944*, 2018.

[56] Qian Wang, Baolin Zheng, Qi Li, Chao Shen, and Zhongjie Ba. Towards query-efficient adversarial attacks against automatic speech recognition systems. *IEEE Transactions on Information Forensics and Security*, 16:896 – 908, 2021.

[57] Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *International conference on machine learning*, pages 5231–5240. PMLR, 2019.

[58] Piotr Zelasko, Sonal Joshi, Yiwen Shao, Jesus Villalba, Jan Trmal, Najim Dehak, and Sanjeev Khudanpur. Adversarial attacks and defenses for speech recognition systems. *arXiv:2103.17122*, 2021.

[59] SM-Moosavi Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. *arXiv:1511.04599*, 2015.

[60] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv:1706.06083*, 2017.

[61] Krishan Rajaratnam, Kunal Shah, and Jugal Kalita. Isolated and ensemble audio preprocessing methods for detecting adversarial examples against automatic speech recognition. *arXiv:1809.04397*, 2018.

[62] Saeid Samizade, Zheng-Hua Tan, Chao Shen, and Xiaohong Guan. Adversarial example detection by classification for deep speech recognition. *arXiv:1910.10013*, 2019.

[63] Krishan Rajaratnam and Jugal Kalita. Noise flooding for detecting audio adversarial examples against automatic speech recognition. *arXiv:1904.12406*, 2018.

[64] Pingchuan Ma, Stavros Petridis, and Maja Pantic. Detecting adversarial attacks on audiovisual speech recognition. *arXiv:1912.08639*, 2021.

[65] Sining Sun, Ching-Feng Yeh, Mari Ostendorf, Mei-Yuh Hwang, and Lei Xie. Training augmentation with adversarial examples for robust speech recognition. *arXiv:1806.02782*, 2018.

[66] Mohammad Esmaeilpour, Patrick Cardinal, and Alessandro Koerich. A robust approach for securing audio classification against adversarial attacks. *arXiv:1904.10990*, 2019.

[67] Keiichi Tamura, Akitada Omagari, and Shuichi Hashida. Novel defense method against audio adversarial example for speech-to-text transcription neural networks. *2019 IEEE 11th International Workshop on Computational Intelligence and Applications (IWCIA)*, 2019.

[68] Sonal Joshi, JesÃºs Villalba, Piotr Zelasko, Laureano-Moro VelÃ¡zquez, and Dehak Najim. Study of pre-processing defenses against adversarial attacks on state-of-the-art speaker recognition systems. *arXiv:2101.08909*, 2021.

[69] Siddique Latif, Rajib Rana, and Junaid Qadir. Adversarial machine learning and speech emotion recognition: Utilizing generative adversarial networks for robustness. *arXiv:1811.11402*, 2018.