Speech Recognition

by Akash Yadav

Submission date: 05-Dec-2023 12:45PM (UTC+0530)

Submission ID: 2248551557

File name: Speech_Recognition.txt (22.68K)

Word count: 3148

Character count: 19657

Speech Recognition

Using

Deep Neural Networks

Submitted By: Akash Kumar Yadav

Course: MCA

Cohort: Batch 3rd

Roll No.: R271223114

SAP ID: 500124804

Submitted To: Anushree Sah

Assistant Professor, SoCS

Date of Submission: 1st November 2023

Reference Paper: Speech Recognition Using Deep Neural Networks: A Systematic Review

Source: https://ieeexplore.ieee.org/document/8632885

INTRODUCTION

Deep learning has emerged as an attractive subject in the realm of machine learning. Its pervasive presence has helped diverse research domains. Deep learning, an cross-disciple of multiple machine learning algorithms, operates through a cascade of layered models. These models predominantly encompass neural networks comprising numerous tiers of non-linear operations. The essence of deep learning lies in the pursuit of gleaning specific features and information from these intricate neural networks.

As mentioned in the article, before 2006, delving into deep architecture inputs was a challenge. Nevertheless, the development of deep learning algorithms has served as a panacea, simplifying the expedition through the parameter space of deep architectures. Deep learning models can also function as a layer wise unsupervised pre-training mechanism, showing a proclivity for learning hierarchies by sequentially extracting features from each stratum. This feature learning paradigm leverages unsupervised learning algorithms, which harness features extracted from the previous layer to prime the subsequent layer. Thus, it endeavors to fathom the transformation of previously acquired features at each subsequent layer. Each iteration of feature learning endows the deep neural network with an additional layer of weights, ultimately serving as a foundation for a deep supervised predictor.

The application of deep architectures has been validated as a more efficient means of representing non-linear functions when compared to shallower architectures. Research has illuminated that deeper architectures require fewer parameters to represent a specific non-linear function than shallower ones. This statistical perspective underscores the efficacy of deeper architectures.

Deep learning algorithms have found prominent utilization in augmenting computer capabilities to mimic human cognition, particularly in domains like speech recognition. Speech, as the primary mode of human communication, has commanded significant attention over the past five decades since the inception of artificial intelligence. Hence, it is only natural that one of the earliest frontiers where deep learning gained traction was in the domain of speech. To this day, an extensive body of research has been published concerning the deployment of deep learning in speech-related applications, particularly in the realm of speech recognition.

Traditional speech recognition systems hinge on the utilization of Gaussian Mixture Models (GMMs) rooted in Hidden Markov Models (HMMs) to represent speech signals. This approach stems from the notion that speech signals can be viewed as piecewise stationary or, in other words, short-time stationary signals. At this shorter timescale, speech signals can be approximated as stationary processes, akin to Markov models characterizing various stochastic processes. Each HMM employs a Gaussian mixture to model a spectral representation of the sound wave. This design is lauded for its simplicity and practicality. Nonetheless, it is deemed statistically inefficient when it comes to modeling non-linear or near non-linear functions.

In stark contrast to HMMs, neural networks enable discriminative training in a much more efficient manner. However, their prowess is most pronounced with short-time signals, such as isolated words. When it comes to continuous speech signals, neural networks often fall short due to their incapability of modeling temporal dependencies. Consequently, a solution arises in employing neural networks as a preprocessing tool, engaging in tasks such as feature transformation and dimensionality reduction for HMM-based recognition. Numerous examples substantiate that the use of deep neural networks yields superior results compared to classical models. A striking instance occurred in 2012 when Microsoft introduced the latest iteration of their Microsoft Audio Video Indexing Service (MAVIS), a speech system built on deep learning. Their final results were resounding, revealing a 30% reduction in Word Error Rate (WER) on four major benchmarks in comparison to state-of-the-art models based on Gaussian mixtures.

This systematic literature review (SLR) adheres to the guidelines outlined by Kitchenham and Charters, with its primary focus on identifying research papers published between 2006 and 2018 in the domain of deep neural networks applied to speech-related applications. These applications encompass a wide spectrum, including automatic speech recognition, emotional speech recognition, speaker identification, and speech enhancement, among others. Initially, 230 papers were identified, but after applying inclusion and exclusion criteria, only 174 papers were deemed suitable for the study. The research questions posed were addressed through a meticulous extraction of information from these 174 papers, culminating in the creation of a statistical representation utilizing tables and figures. The ensuing results intend to elucidate the trajectory of research in this domain over the past years and shed light on emerging research avenues.

Section 2 of this paper provides a concise summary of related work, while Section 3 elucidates the background, encompassing speech recognition and deep neural networks. Section 4 offers insights into the methodology employed in this review, and

Section 5 presents the findings. Finally, Section 6 brings the paper to a thoughtful conclusion, encapsulating the essence of the review's insights and implications.

CONTRIBUTIONS

1. The article provides a valuable contribution by revealing that Mel-Frequency Cepstral Coefficients (MFCCs) are the most commonly used feature extraction method in the analyzed papers. This information can guide researchers in selecting appropriate feature extraction techniques or exploring alternatives to MFCCs.
2. The article contributes by highlighting the use of multiple languages, beyond English, in a subset of research papers. This demonstrates the diversity of language applications in the field of speech recognition, potentially inspiring further cross-lingual research.
3. The article contributes by pointing out the limited research on speech recognition using Recurrent Neural Networks (RNN), particularly Long Short Time Memory (LSTM). This recognition of a research gap can inspire researchers to explore the untapped potential of RNN models in improving speech recognition systems.
4. The article contributes by systematically categorizing the selected research papers into different types, such as conference papers, journal papers, workshop

papers, and research institute publications. This classification provides insights into the sources of research in speech recognition using deep learning.

5. The article's contribution lies in analyzing the distribution of research areas within speech recognition. It identifies the dominant areas, such as speech recognition, speech enhancement, and speaker identification. This analysis helps researchers understand the focus of the field.

6. The article provides insights into the types of deep neural network models used in speech recognition research. It distinguishes between standalone DNN models and hybrid models. This classification can guide researchers in selecting the most appropriate model architecture.

7. The article highlights the limited usage of Recurrent Neural Networks (RNN) in speech recognition research. It suggests that RNNs, particularly Long Short Time Memory (LSTM), have significant potential for improving speech recognition. This insight can motivate future research into exploring the capabilities of RNNs in this field.

MOTIVATION

1. The primary motivation is to provide a comprehensive survey of the application of deep learning techniques in the domain of speech recognition.

2. It seeks to understand the research patterns and preferences in terms of models, techniques, data, and evaluation methods employed in speech recognition research.

3. The article aims to highlight gaps in the existing research, such as the underutilization of recurrent neural networks in speech recognition, which can motivate future research directions.

4. By identifying that [6] Mel-Frequency Cepstral Coefficients (MFCCs) are widely used for feature extraction, the article encourages researchers to explore alternative feature extraction methods to improve speech recognition.

5. The identification of papers using languages other than English motivates research into cross-lingual and multilingual speech recognition, expanding [18] the application of deep learning models to a global context.

PROBLEM

The problem identified is the need for a comprehensive understanding and assessment of the application of deep neural networks in the domain of speech recognition. This problem encompasses several key aspects:

1. Lack of comprehensive survey and analysis of the research conducted at the intersection of deep learning and speech recognition.
2. Underutilization of Recurrent Neural Networks
3. There is a need to explore alternative feature extraction methods beyond the conventional use of Mel-Frequency Cepstral Coefficients (MFCCs) to potentially improve the accuracy of speech recognition models.

OBJECTIVES

1. Classify Speech Recognition Areas
2. Evaluate Performance Metrics
3. Analyze Data Sources
4. Examine Languages and Environments
5. Explore Feature Extraction Methods

PROPOSED METHOD

1. Types of Deep Neural Networks: The article discusses the use of standalone DNN models and hybrid models in speech recognition.

2. Evaluation Techniques: It analyzes the various evaluation techniques used in the research papers, with a focus on metrics like Word Error Rate (WER).

3. Feature Extraction: The article highlights the predominant use of Mel-frequency cepstrum coefficients (MFCCs) as the primary feature extraction method in deep learning models for speech recognition.

ADVANTAGES

1. Improved Recognition Accuracy: Deep neural networks have been shown to outperform traditional models in speech recognition tasks, leading to higher accuracy and better overall performance.

2. Robustness to Variability: DNNs can handle variations in speech signals, such as different accents, speaking rates, and background noise, making them robust for real-world applications.

3. End-to-End Learning: DNNs can learn feature representations directly from raw audio data, eliminating the need for handcrafted feature engineering, which can be time-consuming and error-prone.

4. Scalability: Deep neural networks can scale to accommodate large vocabularies and diverse language models, making them suitable for a wide range of speech recognition applications.

5. Adaptability: DNN models can adapt to individual speakers or environments through techniques like speaker adaptation, resulting in personalized and context-aware speech recognition systems

DISADVANTAGES

1. Data Dependency: DNNs require large volumes of labeled training data to perform effectively. Gathering and annotating such data can be costly and time-consuming.

2. Complexity: DNN models are complex and require substantial computational resources, including high-performance GPUs or TPUs, making them computationally expensive.

3. Overfitting: DNNs are prone to overfitting, especially when training data is limited. Regularization techniques are often needed to mitigate this issue.

4. Lack of Transparency: DNNs are often considered as "black boxes" due to their complex architectures. It can be challenging to interpret and understand why a model makes specific predictions.

5. Hyperparameter Tuning: DNNs involve numerous hyperparameters that need to be fine-tuned, and selecting the right hyperparameters for optimal performance can be a time-consuming process.

ALGORITHMS

1. Convolutional Neural Networks (CNNs): The use of CNNs, which contain convolutional and pooling layers for feature extraction and data reduction, making them effective for image and speech recognition tasks.

2. Recurrent Neural Networks (RNNs): RNNs, which include techniques like [7] Long Short-Term Memory (LSTM) networks, are discussed as a class of deep networks suitable for sequence data modeling, including speech recognition.

3. Systematic Literature Review (SLR): The article's methodology is based on conducting a Systematic Literature Review to identify, analyze, and synthesize existing research papers [17] in the field of speech recognition using deep neural networks.

4. Quality Assessment Rules (QARs): The authors employ Quality Assessment Rules to evaluate the quality of research papers based on specific criteria, helping filter and select papers for inclusion in the review.

5. Feature Extraction Techniques: The article discusses the use of feature extraction methods, with a notable emphasis on Mel-Frequency Cepstral Coefficients (MFCCs) and their application in deep learning models for speech recognition.

SUMMARY

The article is all about using computers to understand and recognize human speech. You know, like when you talk to your smartphone and it responds? That's called speech recognition. Well, scientists and researchers have been working hard to make this technology even better, and they're using something called deep learning.

Now, what's deep learning, you might ask? It's like teaching computers to think like humans. Just like you learn from experience, computers can learn from data. They do this by using something called neural networks, which are like big, interconnected brains inside the computer. These "brains" help computers understand and process speech.

This article talks about a bunch of different things related to deep learning in speech recognition. Let's break down the main points:

Why?

People want computers to understand speech better. This can help us in many ways, like talking to our devices and having them do things for us. Think about how useful it is when you can just say, "Hey, turn on the lights" to your smart home system, and it does exactly that.

What Did They Study?

The researchers looked at a whole bunch of papers and studies (174 to be exact!) to see what others have been doing in this field. They wanted to learn about different ways people are using deep learning for speech recognition. They checked papers from 2006 to 2018 to see what's been happening over the years.

Different Types of Speech Recognition

There are many ways computers can recognize speech. Some people use speech recognition for things like figuring out who's talking (speaker identification) or understanding the emotions in someone's voice (speech emotion recognition). Others use it for converting speech into written text (speech transcription). Most often, it's used for general speech recognition – understanding what people are saying.

Where Do They Get Data?

To teach computers to understand speech, you need lots of examples. This data can come from different places. Sometimes researchers use public data, which is available to anyone. Other times, they have to create their own data. Most of the time, they used English language data, but there were a few studies that used other languages.

What's the Environment?

When we talk, it's not always in a quiet room. Sometimes there's background noise. This article talks about the environments researchers used – like noisy places and regular, quiet rooms.

Measuring Success

How do you know if the computer understands speech? You need to measure it. The article explains the different ways researchers check if their systems work. They use things like Word Error Rate (WER) and Phone Error Rate (PER) to see how many mistakes the computer makes when understanding speech.

Different Methods

To make the computer understand speech, scientists use different methods. Two common ones are Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). CNNs help break down speech into smaller pieces, while RNNs are good at understanding the sequence of words. Both methods are like tools in a toolbox that scientists use to build their speech recognition systems.

Good and Bad Things

Using deep learning for speech recognition has its good sides. It makes computers understand us better, even with all our different accents and voices. It's super handy for things like voice assistants and transcribing voice into text. But there are also some downsides, like the huge amount of data needed to train computers and the fact that some techniques are still not perfect for understanding all languages or in noisy environments.

Enhancements

The article points out that researchers need to explore more ways to improve speech recognition. They mention using Recurrent Neural Networks more because they're pretty powerful. There's also room for coming up with new methods to make speech recognition more accurate and helpful in different situations.

In a nutshell, this article is like a treasure map for scientists working on speech recognition. It shows what others have done and where we can go in the future to make computers understand us even better. With deep learning, it's like we're teaching computers to understand human speech just like we do. And that's pretty amazing!

Techniques Used: • Neural Networks

- Feature Extraction
- Hybrid Models
- Word Error Rate (WER)
- Phone Error Rate (PER)
- Label Error Rate (LER)

Algorithms Used: • Deep Neural Networks (DNNs)

- Convolutional Neural Networks (CNNS)
- Recurrent Neural Networks (RNNs)
- Hidden Markov Models (HMMS)
- Gaussian Mixture Models (GMMs)

Results: 1. The article doesn't provide specific numerical results or findings in a concise format.

2. It primarily presents a systematic review and analysis of existing research on speech recognition using deep neural networks.

3. The article summarizes

- the types of models and
- techniques used,
- their applications, and

their popularity among researchers in the field of speech recognition.

4. It doesn't present its own experimental results but serves as a comprehensive [11] overview of the state of the art in [11] speech recognition with deep neural networks.

Future Enhancement: 1. Future research could explore alternative feature extraction methods. Researchers may investigate the effectiveness of methods like Linear Predictive Coding (LPC) to improve the accuracy of speech recognition systems.

2. Future research could delve deeper into hybrid models, as they have shown promise in improving speech recognition performance. Researchers may explore different combinations of models to optimize results.

3. Future investigations could focus on the application of RNNs for speech recognition, as these models have demonstrated significant potential in handling sequential data.

4. Future research can expand to explore multilingual and cross-lingual speech recognition. Investigating the challenges and solutions for recognizing speech in

multiple languages can enhance the applicability of speech recognition systems in diverse linguistic contexts.

5. The article reports that 27% of papers focused on noisy environments. Future research can concentrate on developing more robust systems for recognizing speech in noisy conditions. This is particularly important for applications like voice assistants and telecommunication systems.

6. Researchers can work on developing and adopting more sophisticated evaluation techniques beyond Word Error Rate (WER) and Phone Error Rate (PER). New methods may provide a more comprehensive understanding of system performance, potentially leading to better models.

7. Exploring alternative deep learning architectures beyond Deep Neural Networks could be an exciting avenue for future research. Variations such as convolutional and recurrent neural networks might yield unique advantages in speech recognition tasks.

8. Future research can focus on the seamless integration of speech recognition systems with real-world applications. This includes improving the accuracy of voice-activated systems, voice assistants, and automated transcription services.

Limitations: 1. Scope Limitations

2. Quality of Included Papers

3. The article covers research published between 2006 and 2018. This limitation could impact the relevance of the findings to current research.

4. The findings might lack context and in-depth insights into the methodologies and algorithms employed in the papers.

5. The article mainly reports on the use of traditional evaluation metrics, such as Word Error Rate (WER) and Phone Error Rate (PER). These metrics, while widely used, may not capture the full spectrum of performance evaluation for speech recognition systems.

6. Insufficient Exploration of Hybrid Models.

REFERENCES

1. Nonlinear Network Speech Recognition Structure in a Deep Learning Algorithm - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8970943/pdf/CIN2022-6785642.pdf

2. Attention-Inspired Artificial Neural Networks for Speech Processing: A Systematic Review - https://www.mdpi.com/2073-8994/13/2/214

3. Deep Neural Networks for Acoustic Modeling in Speech Recognition - https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/38131.pdf

4. Deep Learning for Environmentally Robust Speech Recognition: An Overview of Recent Developments - https://arxiv.org/pdf/1705.10874.pdf

5. Speech Recognition Using Deep Learning Algorithms - https://cs229.stanford.edu/proj2013/zhang_Speech%20Recognition%20Using%20Deep%20Learning%20Algorithms.pdf

6. Accent based speech recognition A critical overview - https://www.malayajournal.org/index.php/mjm/article/view/1739/1184

7. Speech Recognition Using Deep Neural Networks: a Systematic Review - https://www.academia.edu/48095938/Speech_Recognition_Using_Deep_Neural_Networks_a_Systematic_Review

8. Automatic speech recognition: a survey - https://www.researchgate.net/publication/345710977_Automatic_speech_recognition_a_survey

9. Machine Learning in Automatic Speech Recognition A Survey - https://www.researchgate.net/publication/276351194_Machine_Learning_in_Automatic_Speech_Recognition_A_Survey

10. Spoken Language Recognition: From Fundamentals to Practice - https://www.researchgate.net/publication/260686004_Spoken_Language_Recognition_From_Fundamentals_to_Practice

11. Distant Speaker Recognition An Overview - https://www.researchgate.net/publication/280698238_Distant_Speaker_Recognition_An_Overview

12. Language-independent and language-adaptive acoustic modeling for speech recognition - https://www.cs.cmu.edu/~tanja/Papers/SchultzSpecomOrigPublication.pdf

Speech Recognition

ORIGINALITY REPORT

18%	5%	14%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, Khaled Shaalan.. "Speech Recognition Using Deep Neural Networks: a Systematic Review", IEEE Access, 2019 Publication	7%
2	Submitted to Chandigarh University Student Paper	1%
3	Submitted to B.V. B College of Engineering and Technology, Hubli Student Paper	1%
4	Submitted to Liverpool John Moores University Student Paper	1%
5	António Teixeira. "On the Use of Machine Learning and Syllable Information in European Portuguese Grapheme-Phone Conversion", Lecture Notes in Computer Science, 2006 Publication	1%
6	speechdat.org Internet Source	1%

7	pure.ulster.ac.uk Internet Source	1 %
8	Submitted to Stefan cel Mare University of Suceava Student Paper	1 %
9	Submitted to City University Student Paper	<1 %
10	Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, Khaled Shaalan. "Speech Recognition Using Deep Neural Networks: A Systematic Review", IEEE Access, 2019 Publication	<1 %
11	Geoffrey E. Hinton, Simon Osindero, Yee-Whye Teh. "A Fast Learning Algorithm for Deep Belief Nets", Neural Computation, 2006 Publication	<1 %
12	Submitted to Babes-Bolyai University Student Paper	<1 %
13	on-demand-gtc.gputechconf.com Internet Source	<1 %
14	export.arxiv.org Internet Source	<1 %
15	ijisrt.com Internet Source	<1 %
16	www.mdpi.com Internet Source	<1 %

17	Sepp Hochreiter, Jürgen Schmidhuber. "Long Short-Term Memory", Neural Computation, 1997 Publication	<1 %
18	cloud.tencent.com Internet Source	<1 %
19	www.ijraset.com Internet Source	<1 %

Exclude quotes	Off	Exclude matches	< 2 words
Exclude bibliography	On		