Technical University of Munich

THE MATHEMATICS OF SPEECH RECOGNITION AS A CHALLENGE FOR DATA ANALYSIS AND ITS APPLICATIONS

YEVA GALSTYAN

29 November 2023

Declaration of Authorship

I hereby declare that the essay submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the essay in digital form can be examined for the use of unauthorized aid and in order to determine whether the essay as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future theses submitted. Further rights of reproduction and usage, however, are not granted here. This paper was not previously presented to another examination board and has not been published.

Abstract

In today's rapidly evolving world, automatic speech recognition (ASR) is widely recognized as a signature technological achievement, transforming how humans communicate with computers. These models have witnessed remarkable progress, from their early developments in the 1950s, such as "Audrey," to the utilization of mathematical models and Artificial Intelligence (AI) tools. Nowadays, with the integration of Deep Learning and Neural Network components, the application areas of ASR models have extended to the multi-speaker recognition and Speech Emotion Recognition (SER) models. The impact of ASR on human life and its potential applications in diverse fields highlight this technology's ongoing and future relevancy. This topic aims to highlight the challenges that arise in data analysis due to the ongoing advancements in ASR models.

Overview of ASR models

ASR models allow software to convert human speech into written text. It often needs to be corrected for voice recognition, which is concerned with identifying the unique characteristics of the user's voice. In contrast, ASR converts spoken words into written form (*What Is Speech Recognition? | IBM*, n.d.). The underlying mechanism of the ASR model includes gathering distinctive characteristics from the language model, constructing an acoustic model, aligning it with the language model, and employing algorithms for decoding (Leini & Xiaolei, 2021).

The first ASR model, "Audrey" (Automatic Digit Recognizer), was developed at Bell Laboratories in 1952. Audrey was designed to recognize spoken digits 0 to 9 with 90% accuracy, converting them into processable electrical signals. Shortly after, Carnegie-Mellon University introduced an enhanced ASR model known as the Harpy System, which is able to recognize 1000 words. Harpy combined the most effective attributes of the two preceding models, Harpy-1 and Dragon (Lowerre, 1976).

While these models observed a significant milestone in ASR, their limitations in recognizing diverse words and controlled environment requirements restricted their integration into real-world applications.

In the 1980s, the vocabulary capacity of ASR expanded from a few hundred to several thousand words with the introduction of algorithmic solutions based on mathematical models (*A Brief History of Speech Recognition*, n.d.). The hidden Markov Model (HMM) is one of the first mathematical models introduced to the ASR domain, expressing probability distributions across the observation sequence (GHAHRAMANI, 2001). Due to its ability to analyze speech attributes and generate sound models using the time-series signals of spoken language, HMMs are extensively used in constructing acoustic models (Leini & Xiaolei, 2021).

In the 2000s, Google introduced the Google Voice Search service, enabling users to perform web searches and execute commands using natural language. This introduction marks a considerable change in the ASR field, improving human-computer interaction. However, with an increasing demand for user-independent ASR models, the weaknesses of traditional statistical models and basic pattern-matching algorithms became apparent. Background noise, tone, mood, and intonation changes lead to inaccuracies or unclear pronunciations (Leini & Xiaolei, 2021).

Deep learning (DL) models for ASR models

In recent years, speech processing has particularly transformed by integrating powerful tools. The introduction of deep learning models enabled the obtaining of meaningful attributes from raw speech. Additionally, neural networks(NN) such as deep neural networks (DNNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs) have upgraded the analysis and manipulation of speech signals (Mehrish, 2023). Sophisticated deep neural networks with numerous hidden layers have demonstrated superior performance over HMM-based systems across various ASR applications (Hinton et al., 2012).

While deep learning has shown considerable progress in ASR, it faces challenges, such as the need for vast amounts of labeled data, the interpretability of the models, and the ability to maintain flexibility in various environmental conditions (Mehrish, 2023). Denoising technologies address environmental challenges and eliminate background noise using sophisticated Machine Learning (ML) models (Amandeep Singh Dhanjal & Singh, 2023).

Due to the increasing use of neural networks in ASR models and their need for massive datasets, several companies are collecting user data and making it available for further analysis. In particular, Mozilla has established an open-source platform for speech datasets, allowing individuals to contribute their speech data by reading a sentence or accessing the existing dataset for research purposes. However, obtaining labeled data for a specific natural language is sometimes problematic. Recently, Facebook released wave2vec, an open-source library, using a self-supervised algorithm that enables automatic speech recognition with minimal transcribed data (Amandeep Singh Dhanjal & Singh, 2023; Schneider et al., 2019).

An additional barrier to effectively including DL in ASR models is the presence of language variations and diverse accents. Currently, numerous speech-processing datasets are available for data analysis, providing valuable resources for researchers to enhance the flexibility of ASR models to accents. For instance, the VoxForge dataset provides an extensive collection of English, Spanish, French, German, Russian, and Italian speech accents (Amandeep Singh Dhanjal & Singh, 2023; *Voxforge | TensorFlow Datasets*, n.d.).

ASR applications and future advances

Recent advancements in ASR models, coupled with the ongoing progress in DL and NNs, have extended the applications of these models across various domains in technology and science. In recent years, virtual assistants like Alexa, Google Assistant, and Apple Siri have remarkably changed how humans interact with computers, impacting and improving their lives.

Researchers are particularly interested in two main areas: multi-speaker end-to-end ASR and SER models. The rise of teleconferencing and in-car voice assistants has pushed multi-speaker ASR models into a primary research focus (Yifan et al., 2023). End-to-end models built on fully RNNs have demonstrated effectiveness in multi-speaker speech recognition, performing well in single microphone and multi-microphone input scenarios (Chang et al., 2020). On the other hand, SER models are offering potential benefits in areas like eliminating driver fatigue, aiding medical diagnoses, and gathering student feedback in online education. SER models involve identifying and extracting information from speech well suited for the computational identification and differentiation of emotions (Samaneh Madanian et al., 2023).

Conclusion

ASR models reveal a dynamic evolution marked by integrating DL and NN components. As these complex architectures became integral to ASR, the demand for substantial labeled datasets rose. With their complicated learning processes, NNs require extensive data for training and adaptation. The lack of extensive labeled datasets in diverse natural languages and the influence of human accents, background noise, and environmental challenges have elevated the complexity of advancing research and implementing ASR models. This complexity has initiated an increase in various research domains, including noise isolation models and the integration of self-supervised algorithms.

References

A brief history of speech recognition. (n.d.). Sonix. https://sonix.ai/history-of-speech-recognition#:~:text=1950s%20and%2060s

Amandeep Singh Dhanjal, & Singh, W. (2023). A comprehensive survey on automatic speech recognition using neural networks. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-023-16438-y

Chang, X., Zhang, W., Qian, Y., Le Roux, J., & Watanabe, S. (2020). *END-TO-END MULTI-SPEAKER SPEECH RECOGNITION WITH TRANSFORMER*. https://arxiv.org/pdf/2002.03921.pdf

GHAHRAMANI, Z. (2001). AN INTRODUCTION TO HIDDEN MARKOV MODELS AND BAYESIAN NETWORKS. *International Journal of Pattern Recognition and Artificial Intelligence*, *15*(01), 9–42. https://doi.org/10.1142/s0218001401000836

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-R., Jaitly, N., Senior, A., & Vanhoucke, V. (2012). *Deep Neural Networks for Acoustic Modeling in Speech Recognition*. https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/38131.pdf

Leini, Z., & Xiaolei, S. (2021). Study on Speech Recognition Method of Artificial Intelligence Deep Learning. *Journal of Physics: Conference Series*, *1754*(1), 012183. https://doi.org/10.1088/1742-6596/1754/1/012183

Lowerre, B. T. (1976). *LOWERRE, BRUCE T. THE HARPY SPEECH RECOGNITION SYSTEM dtontord Univonrty Libwnes (i Dapt of SpacM Cflfescfcons*. https://stacks.stanford.edu/file/druid:rq916rn6924/rq916rn6924.pdf

Mehrish, A. (2023). *A Review of Deep Learning Techniques for Speech Processing*.

https://arxiv.org/pdf/2305.00359.pdf

Samaneh Madanian, Chen, T., Adeleye, O., Templeton, J. Y., Poellabauer, C., Parry, D., & Schneider, S. (2023). Speech emotion recognition using machine learning — A systematic review. *Intelligent Systems with Applications*, *20*, 200266–200266. https://doi.org/10.1016/j.iswa.2023.200266

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). *WAV2VEC: UNSUPERVISED PRE-TRAINING FOR SPEECH RECOGNITION*. https://arxiv.org/pdf/1904.05862.pdf

voxforge | TensorFlow Datasets. (n.d.). TensorFlow. Retrieved November 30, 2023, from https://www.tensorflow.org/datasets/catalog/voxforge

What is Speech Recognition? | IBM. (n.d.). Www.ibm.com. https://www.ibm.com/topics/speech-recognition

Yifan, G., Yao, T., Hongbin, S., & Yulong, W. (2023). Multi-channel multi-speaker transformer for speech recognition. *INTERSPEECH 2023*. https://doi.org/10.21437/interspeech.2023-257