Speech Recognition

Introduction

The recent acceleration of developments, revolution of technology, economy, culture, lifestyle and social interactions have been only possible with easy global connectivity. Globalization has made the world more interconnected than ever before. People from different countries collaborating has become a norm. This globalized world required effective communication for which ASR technology played a crucial role.

Automatic Speech Recognition (ASR) or voice recognition is a technology that converts spoken words or phrases into a digital format understood by computers. This digital data can then be used for various purposes like interpretation, real time translation, transcription, thus facilitating easy and seamless communication.Since the introduction of ASR it has impacted on various sectors, including tourism, business, diplomacy, and academia, increasing cooperation and reducing linguistic boundaries.

ASR technology has improved significantly in recent years, however it still lacks at places. ASR works well with standard accents, such as British or American English, but the non-standard accents, including African accents are poorly recognized. African accents show unique phonetic and linguistic characteristics which are difficult for current ASR technology to identify. This Paper aims to connect this bridge and make sure that ASR systems are accessible and effective for individuals with diverse linguistic backgrounds across the African continent.

There are over 2,000 distinct languages spoken across Africa in 54 different countries. Along with this diversity there are several regional accents, dialects, and language variations. By developing an ASR system to recognize African accents, we can improve communication, allow easy access to information, and promote modern technical empowerment for millions of peoples of Africa.

This Paper includes exploration of different African accents.The author aims to analyze the phonetic and linguistic characteristics of multiple African accents to get important insights into their uniqueness and the various hardships they present for ASR models.

Next, we seek to evaluate the current performance of ASR systems on African-accented speech and check for areas or gaps which need to be improved. This evaluation will allow us to mark the limitations of our current ASR and determine the challenges faced when recognizing African accents.

Furthermore, this paper aims to study new Deep Learning Algorithms such as Long Short Term Memory(LSTM), convolutional neural networks (CNNs) and some transformer based models. Deep Learning is the field of study in which machines are trained to extract different features after analyzing data using multiple layers of processing known as artificial neural networks

(ANN). ANN are computing systems inspired from human neural networks. These cutting-edge techniques are very effective, and we will use their capabilities to improve african accent recognition.

The Analysis and exploration will help us develop and evaluate a customized ASR model made to handle African-accented speech. In order to measure the effectiveness of our customized ASR model, we will conduct a comparison study of our model with existing ASR systems, the evaluation metrics being the accuracy and recognition rate. Finally, sharing our findings, insights, and methodologies will help guide future researchers and developers to integrate african accented speech recognition in ASR models.

Literature Review

Shibano et al., [1] demonstrated the potential of developing accent-independent and accent-dependent models in order to improve non-native speech recognition. The authors fine-tuned the pre-trained wav2vec 2.0 model using a small labeled dataset and observed better performance for L2 English speakers. The multi-accent and single-accent models showed benefits.

Oladipo et al., [2] deployed a supervised learning algorithm which could recognize accents from three different nigerian ethnic groups, namely Yoruba, Igbo and Hausa using sequential Mel-Frequency Cepstral Coefficients(MFCC) features from the audio samples.

Chan et al., [3] presented Listen, Attend and Spell (LAS), a neural network architecture that enables direct transcription of acoustic signals into characters. LAS uses the sequence-to-sequence framework, incorporating a pyramid structure in the encoder. This structure reduces the decoder's required attention to a smaller number of timesteps, enhancing efficiency and performance.

Li et al., [4] proposed a comprehensive overview of end-to-end (E2E) models and their practical applications. The authors have discussed how E2E models have not only outperformed hybrid models in academic tasks but also have the potential to replace hybrid models in industrial settings.

Maison et al., [5] explored multiple approaches to merge training sets which resulted in significant improvements in performance compared to single-domain baselines. The authors recommend collecting and evaluating additional accents while also investigating novel methods for generating accented speech, particularly for low-resource accents.

Nassif et al., [6] conducted a statistical analysis on the application of deep learning in speech-related tasks. The evaluation techniques employed by the majority of the papers were examined, revealing that the WER (word error rate) was commonly used as a metric to assess the performance and effectiveness of the systems.

Malik et al.,[7] proposed a comprehensive examination and review of various techniques and approaches employed in speech recognition were presented. The analysis of the fundamental architecture of an Automatic Speech Recognition (ASR) system led to the conclusion that three crucial modules—namely, the feature extraction module, classification module, and language model—are integral components of an ASR, shaping its overall performance and functionality.

Lokesh et al., [8] performed an experimental analysis that revealed that the proposed approach achieved an accuracy of 93.6%, outperforming current models such as RNN and DNN-HMM while also exhibiting improved signal-to-noise ratio (SNR) and reduced mean squared error (MSE).

Zue et al.,[9] examined various characteristics of information present in speech signals and highlights the significance of speech-specific knowledge in the advancement of speech recognition systems.

Shaughnessy et al., [10] provided a comprehensive examination of the problem domain, encompassing its methods, achievements, and shortcomings. The author focuses on the characteristics of the speech signal and explores techniques for effective data reduction.

References

[1] Shibano, T., Zhang, X., Li, M.T., Cho, H., Sullivan, P. and Abdul-Mageed, M., 2021. Speech technology for everyone: Automatic speech recognition for non-native english with transfer learning. *arXiv preprint arXiv:2110.00678*.

[2] Oladipo, F.O., Habeeb, R.A., Musa, A.E., Umezuruike, C. and Adeiza, O.A., 2021. Automatic Speech Recognition and Accent Identification of Ethnically Diverse Nigerian English Speakers.

[3] Chan, W., Jaitly, N., Le, Q.V. and Vinyals, O., 2015. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*.

[4] Li, J., 2022. Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, *11*(1).

[5] Maison, L. and Esteve, Y., 2023, June. Improving Accented Speech Recognition with Multi-Domain Training. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.

[6] Nassif, A.B., Shahin, I., Attili, I., Azzeh, M. and Shaalan, K., 2019. Speech recognition using deep neural networks: A systematic review. *IEEE access*, *7*, pp.19143-19165.

[7] Malik, M., Malik, M.K., Mehmood, K. and Makhdoom, I., 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, *80*, pp.9411-9457.

[8] Lokesh, S., Malarvizhi Kumar, P., Ramya Devi, M., Parthasarathy, P. and Gokulnath, C., 2019. An automatic tamil speech recognition system by using bidirectional recurrent neural network with self-organizing map. *Neural Computing and Applications*, *31*, pp.1521-1531.

[9] Zue, V.W., 1985. The use of speech knowledge in automatic speech recognition. *Proceedings of the IEEE*, *73*(11), pp.1602-1615.

[10] O'Shaughnessy, D., 2008. Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, *41*(10), pp.2965-2979.