Comparative Analysis of Automatic Speech Recognition Techniques

Suvarnsing G. Bhable[1(✉)], Ratnadeep R. Deshmukh[1], and Charansing N. Kayte[2]

[1] Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India
suvarnsingbhable@gmail.com, rrdeshmukh.csit@bamu.ac.in
[2] Government Institute of Forensic Science, Aurangabad, Maharashtra, India

Abstract. It's most crucial method of transferring data is interaction. Speech is the most common way of data exchange. According to with linguistic survey, there are 179 languages and 544 dialects spoken in India. Current India has 18 scheduled dialects and several unscheduled languages. The primary goal of this paper is to give a thorough comparative evaluation of the relevant research on automated speech recognition. We observe potential prospects, problems, and methodologies, as well as locate, evaluate, and synthesize data from research in order to give empirical responses to scientific concerns. The survey was done by using appropriate research publications period between 2010 and 2021. The goal of this comprehensive examination is to synthesize the current best research on automated speech recognition by combining the findings of several investigations.

Keywords: ASR · MFCC · DTW · HMM · ML

1 Introduction

Speech recognition is an interdisciplinary area of natural language processing (NLP) that allows computer transcription and conversion through speech into text [1].

Machine learning and Artificial Intelligence are so common and useful in today's society that most people utilize them without thinking about them. The domain of Automatic speech recognition technology becomes a major area in which these modern systems have improved dramatically, nearly to the position in which they are equivalent to human skills [2].

It is regarded as an essential link in developing improved human interaction. The structure of such ASR was made up of the following critical elements: signal processing and feature extract, acoustic modeling, a language model, and hypothesis search [3]. ASR is a slightly elevated process in which a computer converts a voice signal into the relevant text or command after recognizing and interpreting it. ASR entails both extraction and assessment of the acoustic features, the acoustic modeling, as well as the language model. The gathering and evaluation of acoustic features is an important aspect of speech recognition. The retrieval and identification of both the acoustic feature is both data compaction and signals deconvolution method [4] (Fig. 1).

S. Tamane et al. (Eds.): ICAMIDA 2022, ACSR 105, pp. 897–904, 2023.
https://doi.org/10.2991/978-94-6463-136-4_79

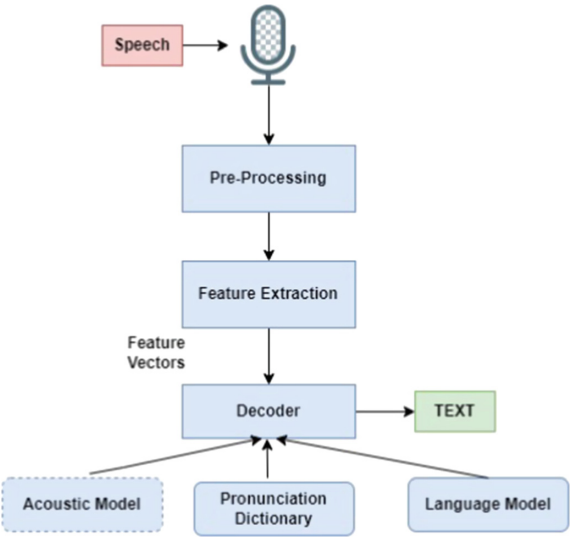

Fig. 1. Automatic Speech Recognition

DNNs are the cornerstone of modern ASR. Acoustic modeling and language model-based interpretation are two components of Speech recognition systems. Hybrid methods often employ DNN to predict the state of every timespan of input, such as assent or states, and hidden Markov models to decode that state information to ultimate transcripts [5].

In this study, a comprehensive and critical review of current strategies is offered with the objective of optimizing how various techniques approach the area of Multilingual Automatic Speech Recognition, which might provide significant perspective in future studies.

2 Related Work

In this study, they reported our finalized recognize systems for such Verbmobil challenge, which were constructed for something like the three different languages German, English, and Japanese. When integrated with both the language identification element, the user is supplied with a versatile and user-friendly multilingual spoken conversation system [6].

In this research, they improve multilingual ASR effectiveness in 2 directions: 1) by analyzing the influence of feeding with the one vector designating the language, and 2) by designing the problem with a meta-learning aim mixed with self-supervised learning (SSL) [7].

An E2E multilingual platform that is prepared to work in limited separating active applications while also addressing a significant difficulty of actual data: an unbalance in the testing phase throughout languages. Utilizing 9 Indian languages, we examine several strategies and discover that a combo of retraining on the language vector and training language-specific adaptor layers gives the best result. The resultant E2E multilingual

model has a decreased word error rate (WER) than either of monolingual E2E models (eight of nine languages) and monolingual local substations [8].

The development of the multi-layer perceptron neural paradigm marks the start of the machine learning revolution. Speech recognition technology has indeed been studied and developed in greater depth, and the advent of artificial neural networks and the integration of models has propelled speech recognition technology to a new level. Used English verb phrases as the object and achieved verb phrase recognition by lexical annotation, named entity recognition, and rule restrictions. Its procedure is quite time-consuming [9]. To assess the effectiveness of Recurrent neural network and completely focused time-delay neural network (FFTDNN) structures in identifying feeling, dialect, speaker, and gender differences in phonemic Assamese languages, this was discovered that machine learning-based sentence extraction technique together with RNN using a composite feature model as a classifier outperformed other techniques in terms of detection accuracy and computationally efficient under several circumstances [10].

Active learning provides the ability to respond to non-stationary events via a feedback mechanism embedded into the training algorithm. Active learning may also be seen as an optimization method that picks training instances to maximize test set word correctness [11].

3 Challenges for Automatic Speech Recognition

3.1 Accuracy

Accuracy refers to more than just the accuracy of the word output – the WER. Many other factors affect the level of accuracy on a case-by-case basis. These factors are often Background noise unique to a use case or a particular business need and include:

- Background noise
- Punctuation location
- Capitalization
- Proper formatting
- Word order
- Domain-specific terminology
- Identifying the speaker

3.2 Deployment

To eliminate this hurdle to acceptance, providers should make their deployments and integrations as painless as possible.

3.3 Language Coverage

Many of the main speech technology vendors have language coverage gaps. Most providers offer English, but when multinational enterprises wish to employ speech technology, a lack of language coverage creates a hurdle to adoption. Even when providers do provide additional languages, accent or dialect identification is often a challenge.

3.4 Data Security and Privacy

Data security is under threat as a result of the media's depiction of 'data-hungry' digital behemoths. It might also be the consequence of more day-to-day talks taking place online as a result of the coronavirus epidemic, which increased remote working.

4 Research Methods for Speech Recognition

4.1 Neural Network Based Speech Recognition

In this section, we discussed several previous studies linked with the neural network-based voice recognition approach that encouraged us to do this study. In this section, several of the studies are clarified briefly. The Parallel Implementation of Artificial Neural Network Training for Speech Recognition has been approved. They demonstrated the execution of a full ANN training procedure utilizing the block mode back-propagation learning algorithm for sequential patterns such as the observation feature vectors of a speech recognition system utilizing the high-performance SIMD architecture of GPU using CUDA and its C-like language interface [12].

4.2 Fuzzy Logic Based on Speech Recognition

Instead of detecting and removing noise, fuzzy modeling and decision making ignore it. The speech spectrogram was converted into a fuzzy linguistic explanation for this purpose, and this explanation was used instead of precise acoustic data. During the guiding stage, a genetic algorithm identifies appropriate definitions for phonemes, and after these definitions are defined, a simple new operator including low-cost functions such as Max, Min, and Average formulates the recognition [13].

4.3 Wavelet Based Speech Recognition

The inadequacies of the Fourier transform prompted the development of wavelet transforms. When the FT depicts a signal in the frequency domain, it cannot distinguish where those frequency components are in time. Cutting the signal at a specific point in time (windowing) and transforming it into the frequency domain to obtain a relevant time sequence of frequency information is equivalent to convolving the signal and the cutting window, which may result in smearing of frequency components along the frequency axis [14].

4.4 DTW Algorithm-Based Speech Recognition

DTW has been used to compare various speech patterns in automated speech recognition and to identify an ideal alignment between two provided sequences under specific constraints. Intuitively, the sequences are distorted in a nonlinear form to match each other. DTW has been successfully used to automatically deal with temporal deformations and varied speeds related with time-dependent data [15].

4.5 Sub-band Based Speech Recognition

Low-pass and high-pass filters with various cutoffs were used to investigate individuals' capacity to accurately distinguish phonemes with restricted frequency bandwidths. To eliminate any linguistic knowledge, the participants listened to nonsensical syllables. He discovered that the phoneme identification error rate for a certain band was equal to the product of the error rates for the component sub-bands [16].

5 Challenges in Speech Recognition

- Accurately and efficiently convert a speech signal into a text message independent of the device, speaker or the environment.
- Acoustic Training Issues
- Language Model Training Issues
- Automatic creation of word lexicons.
- Language models for new jobs are generated automatically.
- Identifying the theoretical limit for the deployment of automated voice recognition.
- The best utterance verification-rejection algorithm.
- Achieving or exceeding human performance on ASR tasks

6 Toolkits for ASR

- AT&T FSM Library
- CMU-Cambridge Statistical LM Toolkit
- CMU Sphinx
- CSLU toolkit
- Edinburgh Speech Tools Library
- KTH WaveSurfer
- MSState ASR Toolkit, NIST Utility Software
- SPRACHcore software package
- SRI Language Modelling Toolkit
- Transcriber
- HTK
- Kaldi
- Whisper OpenAI

Table 1. Multilingual Automatic Speech Recognition System in Indian Scenario

Sr. No.	Database	Year	Technique used	Languages	Result	Ref
1	TIMIT	2004	MFCC, HMM	Hindi and Tamil	90.03%	[17]
2	PSTN, NTIMIT	2005	MFCC, HMM	Tamil, Hindi & American English	T-96.89% H-95.31%	[18]
3	TIMIT, WSJ	2018	TDNN, GMM-HMM DNN-HMM	Tamil, Telugu & Gujarati	WER TA-16.35%, TE-18.61% G-12.7%	[19]
4	Train-1500 h Test-90 h	2018	MFCC, LAS	Bengal, Gujarati, Hindi, Kannada, Malayalam, Marathi, Tamil, Telugu, Urdu	WER 16.8, 18.0, 14.4, 34.5, 36.9, 27.6, 10.7, 22.5, 26.8.	[20]
5	YouTube	2018	MFCC, Ensemble learning-based classification	English, Hindi & Bangla	E+ H-99.52% B- 8.70%	[21]
7	Manually Collected Data	2021	MFCC, Context-independent (CI) (GMM)-HMM), GMM-UBM, Baum-Welch (BW)	Kannada,Telugu, Bengal,Odia, Urdu, & Assamese	Multi-PRS 31.5	[22]
8	Microsoft Transcribed collected Data	2022	MFCC, HMM, GMM, DNN	Tamil, Telugu, & Gujarati	WER 9.66	[23]

7 Conclusion

Speech is the main and most convenient form of communication between individuals. Building automated systems that can interpret spoken language and recognize speech in the same way that humans can is a difficult undertaking. The purpose of automatic speech recognition research is to address the many methodologies related to ASR Various approaches have been found and applied to the ASR system, Which Database used, resulting in many successful ASR applications in restricted domains. Some of the study topics gaining traction include robust speech recognition, multimodal speech recognition, and multilingual speech recognition. In the future, we want to work on multilingual automatic speech recognition for Indian languages.

References

1. Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., & Ringel Morris, M. (2019, October). Sign language recognition, generation, and translation: An interdisciplinary perspective. In The 21st international ACM SIGACCESS conference on computers and accessibility (pp. 16–31).
2. Fadhil, A. (2018). Can a chatbot determine my diet?: Addressing challenges of chatbot application for meal recommendation. arXiv preprint arXiv:1802.09100.
3. Yu, D., & Deng, L. (2016). Automatic speech recognition (Vol. 1). Berlin: Springer.
4. Mustafa, M. B., Salim, S. S., Mohamed, N., Al-Qatab, B., & Siong, C. E. (2014). Severity-based adaptation with limited data for ASR to aid dysarthric speakers. PloS one, 9(1), e86285.
5. Lüscher, C., Beck, E., Irie, K., Kitza, M., Michel, W., Zeyer, A., ... & Ney, H. (2019). RWTH ASR Systems for LibriSpeech: Hybrid vs Attention--w/o Data Augmentation. arXiv preprint arXiv:1905.03072.
6. Koehn, P. (2020). Neural machine translation. Cambridge University Press.
7. Ma, C. Y. (2001). Multilingual speech recognition and its application in a multilingual voice browser (Doctoral dissertation).
8. Diwan, A., Vaideeswaran, R., Shah, S., Singh, A., Raghavan, S., Khare, S., ... & Abraham, B. (2021). Multilingual and code-switching ASR challenges for low resource Indian languages. arXiv preprint arXiv:2104.00235.
9. Schmidt, J., Marques, M. R., Botti, S., & Marques, M. A. (2019). Recent advances and applications of machine learning in solid-state materials science. npj Computational Materials, 5(1), 1–36.
10. Pandey, S. K., Shekhawat, H. S., & Prasanna, S. R. M. (2019, April). Deep learning techniques for speech emotion recognition: A review. In 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA) (pp. 1–6). IEEE.
11. Bisong, E. (2019). Building machine learning and deep learning models on Google cloud platform: A comprehensive guide for beginners. Apress.
12. Gill, N. S. (2019). Artificial neural networks applications and algorithms. Dosegljivo: https://www.xenonstack.com/blog/artificial-neural-networkapplications/. [Dostopano: 1. 9. 2019].
13. Abbaschian, B. J., Sierra-Sosa, D., & Elmaghraby, A. (2021). Deep learning techniques for speech emotion recognition, from databases to models. Sensors, 21(4), 1249.
14. Wirsing, K. (2020). Time Frequency Analysis of Wavelet and Fourier Transform. In Wavelet Theory. London, UK: IntechOpen.
15. Nair, N. U., & Sreenivas, T. V. (2008, November). Multi pattern dynamic time warping for automatic speech recognition. In TENCON 2008–2008 IEEE Region 10 Conference (pp. 1–6). IEEE.
16. Vickers, D. A., Moore, B. C., & Baer, T. (2001). Effects of low-pass filtering on the intelligibility of speech in quiet for people with and without dead regions at high frequencies. The Journal of the Acoustical Society of America, 110(2), 1164-1175.
17. Udhaykumar, N., Ramakrishnan, S. K., & Swaminathan, R. (2004). Multilingual speech recognition for information retrieval in Indian context. In Proceedings of the Student Research Workshop at HLT-NAACL 2004 (pp. 1–6).
18. Kumar, C. S., Mohandas, V. P., & Li, H. (2005). Multilingual speech recognition: A unified approach. In Ninth European Conference on Speech Communication and Technology.
19. Fathima, N., Patel, T., Mahima, C., & Iyengar, A. (2018, September). TDNN-based Multilingual Speech Recognition System for Low Resource Indian Languages. In INTERSPEECH (pp. 3197–3201).

20. Toshniwal, S., Sainath, T. N., Weiss, R. J., Li, B., Moreno, P., Weinstein, E., & Rao, K. (2018, April). Multilingual speech recognition with a single end-to-end model. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4904–4908). IEEE.
21. Mukherjee, H., Dhar, A., Obaidullah, S. M., Santosh, K. C., Phadikar, S., & Roy, K. (2018, November). Identification of top-3 spoken Indian languages: an ensemble learning-based approach. In 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN) (pp. 135–140). IEEE.
22. Manjunath, K. E., Raghavan, S. K., Rao, K. S., Jayagopi, D. B., & Ramasubramanian, V. (2021). Approaches for multilingual phone recognition in code-switched and non-code-switched scenarios using Indian languages. ACM TRANSACTIONS on Asian and low-resource language information processing, 20(4).
23. Madhavaraj, A., & Ganesan, R. A. (2022). Data and knowledge-driven approaches for multilingual training to improve the performance of speech recognition systems of Indian languages. arXiv preprint arXiv:2201.09494.