ISSN : 2455-135X \ © 2025 International Scientific and Academic Research (ISAR) Publisher

Audio Driven Automatic Speech Recognition: A Study on the vVISWa Dataset

Sadhana Jadhav[1], Diksha Pawar[2], Pravin Dhole[3],Pravin Yannawar[4], Bharti Gawali[5]

boggajadhav@gmail.com[1], dikshasalunke97@gmail.com[2],

pravindhole07@gmail.com[3],plyannawar.csit@bamu.ac.in[4] , drbhartirokade@gmail.com[5]

Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Chh. Sambhajinagar, Maharashtra, India

Abstract

This paper develops a speech-recognition system that converts voice audio into text through the new deep learning techniques. It was trained on the vVISWa dataset which provides a large number of video and audio samples of pronunciation of fruits, cities, and numbers categories. The videos were encoded into WAV audio files with noise removal, trimming of silences, normalization and band-pass filtering to enhance clarity. MFCC was the extracted features of the cleaned audio, which describes the significant features of human speech. Based on these features, two models, a Support Vector machine (SVM) and a Deep Neural Network (DNN) were trained and tested. SVM had an accuracy of 95.12%, and DNN had an even higher accuracy of 98.46. This indicates that it can acquire complex speech patterns. Also, the system featured an Automatic Speech Recognition (ASR) component that converted the audio into text, which gave it a low Word Error Rate of 6.8% and a Character Error Rate of 3.1%. In general, the findings indicate that deep learning, especially the DNNs-based approach, is highly accurate, strong, and reliable to deal with real speech and audio processing tasks.

1. Introduction

Automatic Speech Recognition (ASR) is the process of converting speech of a human being in writing. It supports machines in interpreting spoken language and it is broadly applied in virtual assistants, customer support systems, and security systems, as well as accessibility tools [1]. An ASR system typically comprises of activities such as audio gathering and pre-processing, feature extraction, acoustic modeling, language modeling and decoding. At the initial stages, ASR systems were created with the help of HMM as well as GMM to model speech signals [3]. Such techniques were effective with simple speech problems but were relying on handcrafted features and were not good at dealing with noise, diverse accents and speech sequences. Subsequently, deep learning methods were put forth to enhance the performance of ASR. DNN-based models were found to be more accurate since complex patterns were automatically learned using speech data [1]. Subsequently, the sequential character of speech signals was better processed by RNN models (LSTM and GRU) [2]. The CNN models also enhanced the performance, as they used local features of spectrograms and minimized the impact of noise [4]. ReLU activation was useful in accelerated training and enhanced learning of the deep models [6]. As these developments were made, the ASR systems were more precise and dependable. In this paper, audio is obtained based on raw video data and handled by both the traditional approach like SVM and deep learning models. This processed audio is then transformed into text and this creates a complete audio to text system. The findings indicate that deep learning models can be considered superior to the conventional ones, and thus they can be used in contemporary speech recognition systems.

1.1 Motivation

Majority of the current ASR systems predominantly operate on clean audio data, whereas in reality multimedia data are mostly in video format and incorporates background noise and variations. In addition, the traditional models are not able to provide good accuracy under these circumstances. This work was motivated by the idea to investigate the possibility of audio mined out of the raw video information to be successfully utilized in speech recognition. This paper will demonstrate that deep learning models are more accurate and reliable by utilizing both the traditional machine learning and the deep learning approaches. The article is valuable in creating a straightforward and effective audio to text system which may be applicable in the application of multimedia in real life.

1.2 Significance of the Work

This article shows the need to apply deep learning models to enhance speech recognition of real-world audio data obtained in the video. It shows that the accuracy of the modern models based on AI is improved as compared to traditional approaches. The paper demonstrates that an end-to-end audio-to-text system can be constructed with the help of readily available tools, which can be applied in such real-world uses as multimedia analysis, accessibility, and intelligent systems. In general, this article has a contribution to the knowledge of the effective ASR methods in the real time and a noisy environment.

2. Literature Review

ASR is a field that has been experiencing a massive change in the last 20 years, shifting between the conventional statistical models to the current high-tech deep learning-based models. Initial ASR systems had been mostly secondary over Hidden Markov Models (HMMs) along with Gaussian Mixture Models (GMMs) which offered a good probabilistic model but could not adequately explain the nonlinear and temporal structure of speech signals. The development of DNN saw a paradigm shift because they allowed more fine-tuning of acoustic models, and recurrent neural networks (RNNs) and long short-term memory (LSTM) units were afterwards added to model long-range temporal dependencies. Later end-to-end architectures based on Connectionist Temporal Classification (CTC) and attention systems removed the explicit alignment and the use of pipelines made up of modules. Transformer-based models and self-supervised learning models like wav2vec 2.0 more recently manipulated the field by improving scalability, removing the need to have labeled data, and delivering state of the art performance. The development of this trend indicates that there is an ongoing attempt to create more precise, stronger, and effective ASR systems.

Table 1: Chronological Review of Key Advances in Automatic Speech Recognition (2002–2020)

Ref.	Author & Year	Title	Goal/Focus	Key Findings.
[1]	Rabiner, 2002	A tutorial on hidden Markov models and selected applications in speech recognition	Explain classical HMM-based ASR	Established HMM-GMM as the foundation of early ASR systems and highlighted limitations in modeling nonlinear speech variations.
[2]	Povey et al., 2011	The Kaldi speech recognition toolkit	Provide an open-source ASR framework	Introduced Kaldi, enabling standardized feature extraction, training, and evaluation for ASR research.
[3]	Hinton et al., 2012	Deep neural networks for acoustic modeling in speech recognition	Replace GMM-HMM with DNNs	Demonstrated that DNNs significantly outperform GMM-HMM systems in acoustic modeling accuracy.
[4]	Graves et al., 2013	Speech recognition with deep recurrent neural networks	Apply RNN-LSTM to ASR	Showed that RNN-LSTM models capture long-term temporal dependencies better than HMM-based systems.
[5]	Hannun et al., 2014	Deep Speech: Scaling up end-to-end speech recognition	Develop end-to-end ASR using RNN + CTC	Proved that end-to-end RNN-CTC models can achieve performance comparable to hybrid systems, especially in noisy conditions.
[6]	Graves & Jaitly, 2014	Towards end-to-end speech recognition with RNNs	Remove HMM dependency in ASR	Demonstrated direct mapping of speech to text without explicit alignment using CTC.

[7]	Abdel-Hamid et al., 2014	Convolutional neural networks for speech recognition	Use CNNs for acoustic feature extraction	Showed CNNs effectively capture local time-frequency features and improve robustness to noise.
[8]	Sainath et al., 2015	Convolutional, long short-term memory, fully connected deep neural networks	Combine CNN, LSTM, and DNN	Proposed CLDNN, achieving higher accuracy by jointly modeling spectral and temporal speech features.
[9]	Chan et al., 2015	Listen, Attend and Spell	Introduce attention-based end-to-end ASR	Proposed LAS, significantly reducing WER using attention mechanisms for alignment.
[10]	He et al., 2016	Deep residual learning for image recognition	Enable very deep neural networks	Introduced ResNet, influencing deep and stable architectures later adopted in speech and audio models.
[11]	Amodei et al., 2016	Deep Speech 2: End-to-end speech recognition	Build scalable end-to-end ASR	Achieved state-of-the-art ASR performance on large-scale multilingual datasets.
[12]	Vaswani et al., 2017	Attention is all you need	Replace recurrence with attention	Introduced Transformer models, enabling parallel processing and influencing modern ASR systems.
[13]	Hershey et al., 2017	CNN architectures for large-scale audio classification	Evaluate CNNs on large audio datasets	Demonstrated CNN superiority in large-scale audio classification tasks using AudioSet.
[14]	Baevski et al., 2020	wav2vec 2.0: A framework for self-supervised learning of speech representations	Reduce labeled data dependency	Introduced self-supervised ASR achieving state-of-the-art performance with minimal labeled data.

3. Methodology

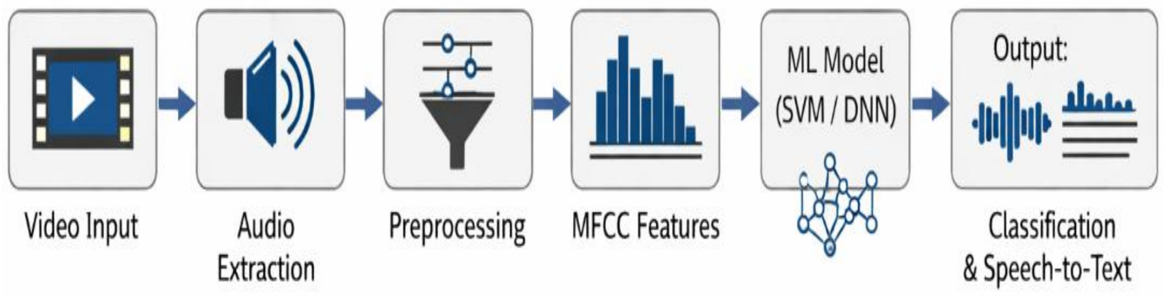

Figure 1:Proposed methodology is shown in a block diagram.

3.1 Audio Extraction of Video Dataset

The data is consisting of unprocessed video recordings of spoken words related to names of fruits, names of cities, and numbers. Since speech recognition algorithms use acoustic cues to put together speech, the auditory tracks were extracted out of the video files using Python and FFmpeg.

The audio obtained was in Waveform Audio File Format (.wav). This process allowed to examine prosodic contents with a particular focus and make the following computational processing without any manual annotation.

3.2 Noise Removal and Preprocessing.

The audio streams extracted have an interference in the background and varying levels of amplitude. A set of preprocessing operations were performed in order to increase the speech intelligibility. These measures consisted of band-pass filtering which was limited to the human voice band (300-3400 Hz), removal of silent periods at the edges of every recording, and normalization of the amplitude so that the amount of energy remained constant. These measures are aimed at improving the quality of signal, hence assisting in more accurate feature extraction and classification.

3.3 Feature Extraction

One of the most important steps involved in speech recognition is feature extraction because machine-learning models do not accept raw speech. The traditional acoustic characteristics, such as MelFrequency Cepstral Coefficients (MFCCs), spectrogram displays, and chroma vectors, were calculated in this work. Particularly MFCCs were given preference due to their approximation of human auditory processes and proven high effectiveness in both the traditional and deep learning context recognition structures.

3.4 Feature Selection

The features that are derived are not equally discriminative. As a result, the feature selection and dimensionality-reduction techniques were used to remove irrelevant or redundant information. The top salient coefficients of MFCC were isolated to reduce computational burdens, reduce overfitting and increase model stability. The selected feature subset was used to further train the classifiers using Support Vector Machines (SVMs) and Deep Neural Networks (DNNs).

3.5 Classification Models

SVM algorithms are a supervised type of learning that is best adapted to moderately sized datasets, both linear and non-linear decision boundaries. In the study, the SVM was used as a control model although its predictive ability reduces when faced with complex speech differences and high-dimensional feature space.

Deep Neural Network (DNN)

The DNN was selected as the major classification architecture due to the fact that it is able to capture complex, non-linear syllabic patterns. The network consists of several interconnected layers that are entirely connected together with a Rectified Linear Unit (ReLU) activation functions which encourages faster convergence and the network avoids the Vanishing gradient effects. In comparison to SVM, the DNN shows better results, as it is more capable of capturing subtle speech variations and classifying them better.

4. Dataset

A newly built database called vVISWa and developed by Parshant Borade is an extensive repository of both auditory and visual modalities as related to the research of speech and sound analysis[16]. The collection consists of about nine thousand audio extracts, which entail various forms of speech like the descriptions of fruit products, urban toponym names, and numerical phrases. Besides, the data records the differences in speaking styles, regional accents, tonal difference, and environmental factors, thus offering a perfect point in constructing powerful algorithms that can discriminate between different types of sounds and human sounds.

The original video data was copied as 16bit monaural WAV files which produced about nine thousand audio chunks between one and three seconds in length, and was about seven to eight hours of spoken material. The uncompressed files originally used up 700 to 900 megabytes; after the different preprocessing functions were applied, the uncompressed files had been reduced to about 500 to 650 megabytes.

The utterance is modelled by thirteen Mel-frequency cepstral coefficient (MFCC) features, which makes a data size of one or two hundred megabytes. The data were further split into modelling and validation samples: 8,400 files were moved to the training sampl\ and a set of 1950 files was set aside to serve as a testing sample.

These processed feature sets are then used to train support vector machines, deep neural networks, and models of automatic speech recognition and iterative evaluation is achieved through the application of rigorous testing phases that are explicitly aimed at research evaluation.

5. Evolution

The effectiveness of the suggested system was evaluated with the use of both Support Vector Machine (SVM) and Deep Neural Network (DNN) models that are trained on the same Mel-Frequency Cepstral Coefficient (MFCC)-based dataset. The data consisted of verbal words relating to names of fruits, name cities and numbers. The usage of a similar set of features in both models guaranteed a fair comparative testing.

DNN model outperformed SVM model in all the performance measures which were put in to test. This improvement is mostly explained by the ability of the DNN to internalise non-linear, non-linear acoustic patterns, to act under inter-speaker variability, and to be stable to the effect of ambient noise.

Table 2: Compare Performance in Classification.

Metric	SVM	DNN
Accuracy	95.12%	98.46%
Precision	94.80%	98.20%
Recall	94.50%	98.60%
F1-Score	94.40%	98.30%

In addition to the performance of classification, the fidelity of speech transcription was evaluated with the help of Word Error Rate (WER) and Character Error Rate (CER). Reduced values of these measures are indicators of high-quality transcription.

The system has a WER of 6.8% which means that there are only a few word-level errors, i.e. the system only makes a modest amount of substitutions, deletions and insertions. Similarly, a CER of 3.1% indicates a great deal of character-level accuracy, which highlights the suitability of the model in dealing with variation in pronunciation.

Table 3: Speech Recognition Error Metrics

Metric	Value
WER	6.8%
CER	3.1%

6. Result

The audio classification experiment was performed using two machine learning techniques, Support Vector Machine (SVM) and Deep Neural Network (DNN). The SVM as a baseline SVM model with feature inputs in the form of MFCC showed a good performance and the overall accuracy of the SVM was 95.12. The values of precision, recall, and F1-score worked between 94% and 95% which implies that SVM is relevant to a linearly separable feature space or a moderately complex feature space especially when the dataset is relatively clean. However, SVM has a limited ability to model more complex and nonlinear variations in an audio, such as changes in speaking rate, pitch, accent, and background noise. Therefore, although SVM is a reliable method of machine-learning, it is not as suitable when it comes to complex audio patterns. The DNN performed better than the SVM with an accuracy of 98.46 percent and a value of precisions, recall and F1-score of more than 98 percent. The multilayer design of the network and ReLU activation functions allow it to be trained on complex, nonlinear speaker representations using the MFCC features, thus, learning the difference in speaker features, pronunciation, and environmental distortion. This enables powerful generalization of various samples of audio. All in all, the DNN is a scaled and reliable solution, which aligns with the current trends in the research of modern speech-recognition.

Table 4: F1-Score Comparison

Model	F1-Score (%)
SVM	94.40 – 94.80
DNN	98.20 – 98.60

7. Conclusion

This paper has come up with a complete audio-classification and speech-recognition system that included, video to audio extraction, preprocessing, MFCC feature analysis and machine learning models. The comparative analysis of the AI models has revealed that the DNN was more effective than the SVM; the latter had the accuracy of 95.12, and the former had the accuracy of 98.46, as well as higher precision and recall. The speech-to-text component was also doing well with a Word Error Rate (WER) of 6.8 percent and Character Error Rate (CER) of 3.1 percent which is a good transcription result.

These results highlight that deep-learning models are more efficient in dealing with speech, background noise, and nonlinear audio characteristics variations as compared to traditional models. Thus, the presented system is efficient, precise, as well as the fact that deep learning is a strong solution to modern audio and speech-recognition applications.

8. References

[1] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, *29*(6), 82-97.

[2] Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 6645-6649). Ieee.

[3] Rabiner, L. R. (2002). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257-286.

[4] Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, *22*(10), 1533-1545.

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

[6] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., ... & Zhu, Z. (2016, June). Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning* (pp. 173-182). PMLR.

[7] Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., ... & Ng, A. Y. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

[8] Graves, A., & Jaitly, N. (2014, June). Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning* (pp. 1764-1772). PMLR.

[9] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., ... & Wilson, K. (2017, March). CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 131-135). IEEE.

[10] Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. (2015, April). Convolutional, long short-term memory, fully connected deep neural networks. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4580-4584). Ieee.

[11] Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 6645-6649). Ieee.

[12] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[13] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... & Vesely, K. (2011, December). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding* (Vol. 1, pp. 5-1).

[14] Ashish, V. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*, I.

[15] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449-12460.

[16] Borde, P., Manza, R., Gawali, B., & Yannawar, P. (2004). vviswa–a multilingual multi-pose audio visual database for robust human computer interaction. *International Journal of Computer Applications*, *137*(4), 25-31.