A Critical Insight into Automatic Visual Speech Recognition System

Kiran Suryavanshi[1](✉), Suvarnsing Bhable[1], and Charansing Kayte[2]

[1] Department of Computer Science & IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, MH, India

[2] Government Institute of Forensic Science, Aurangabad, MH, India

Abstract. This research paper investigated the robustness of the Automatic Visual Speech Recognition System (AVSR), for acoustic models that are based on GMM and DNNs. Most of the recent survey literature is surpassed in this article. Which shows how, over the last 30 years, analysis and product growth on AVSR robustness in noisy acoustic conditions has progressed? There are various categories of languages covered, with coverage, development processes, Corpus, and granularity varying. The key advantage of deep-learning tools, including a deeply convoluted neural network, a bi-directional long-term memory grid, a 3D convolutional neural network, and others, is that they are relatively easy to solve such problems, which are indissolubly linked to feature extraction and complex audiovisual fusion. Its objective is to act as an AVSR representative.

Keywords: Speech recognition · AVSR · MFCC · HMM · CNN · DNN

1 Introduction

Human-computer collaboration is a pursuit of HCI as human-computer communication for the interaction between humans and computers (HCI). Given the rapid advancement and promotion of computer intellectualization in Artificial Intelligence, HCI technology is presently confronted by more challenges and complexity than before. With this in mind, a most effective and humanized audio perception method appears feasible when dealing with large HCI issues, whether the devices run in a nice work surrounding or in a noisy working environment. ASR is an effective link between the two components of humans and machines; ASR (Audio speech recognition) is an effective link [1, 2].

The AVSR may be utilized in many settings, including ground vehicle signal recognition, mobile text translation, lip-reading for persons with hearing impairments, voice identification of speakers out of a lot of persons speaking simultaneously time, etc. [3].

The AVSR's historical period in that we summarized the conventional and deep learning-based methods used in AVSR systems by mathematically representing the AVSR operation. Meanwhile, we compared the existing AVSR dataset with the mission, testers, corpus, perspective (or view), and so on, and explain our ideas in practical scenarios of open-source audiovisual databases [4].

A. Dev et al. (Eds.): AIST 2021, CCIS 1546, pp. 3–11, 2022.
https://doi.org/10.1007/978-3-030-95711-7_1

We focused on two major issues when constructing AVSR anatomy. The function extraction (in terms of audio/visual modality) and complex audiovisual fusion are both thoroughly investigated in theory. Extractors based on convolutional neural networks (CNNs) that allow efficient feature extractions and advanced neural networks [5].

For continuous number digits registered in the TIDIGIT database, the speech recognition rate for the MFCC-based ASR system is higher than the speech recognition rate for the PLP-based ASR system. In the same report, however, PLP-based ASR systems outperformed MFCC-based ASR systems in terms of accurate phoneme accuracies for speech data from the TIMIT database [6].

2 Related Work

The deep neural network hidden Markov model (DNN-HMM) Sanskrit speech recognition using HTK is one of the most promising implementations of DNNs in ASR. For extraction, MFCC and two states of HMM were used, yielding 95.2% and 97.2% precision, respectively [7]. For Hindi words, a real-time speaker recognition system was created. MFCC was used to remove features using the Quantization Linde, Buzond Gray (VQLBG) algorithm. To break the silence, the Voice Activity Detector (VAC) was suggested [8].

Technology is used. Linear Predictive Coding (LPC) and Gaussian Mixture Model were used to remove features (GMM). A total of 100 words were reported 1000 times, yielding an accuracy of 84% [9].

Auto HTK language recognition. Isolated terms are used to identify HMM topology speech in 10 states that generated 96,61% [10]. A Hidden Markov Model Toolkit is an automated voice recognition framework for isolated and associated Hindi words (HTK). Hindi terms are used with MFCC datasets and 95% isolated words and 90% in related words are used with a recognition scheme [11].

A successful speech recognition system experimented with a 98% precision Mel Frequency Cepstrum Coefficients (MFCC), Vector Quantization (VQ) [12], HMM. The index is made up of five words spoken ten times by four voices. We need a sense of how the voice is generated to grasp the relationship between the auditory sound and the associated visual signals which can be perceived on the speaker's mouth/lips. The two areas of grammar that research sounds of human language are phonetics and phonological science. Phonetics investigates the language's actual speech sounds, and how the sound is produced, communicated, and seen. On the other hand, phonology is a formal analysis of the organization of speech sounds to create sound structures. Phonetics is linked to acoustic research by using a great deal of sound processing technically employed by acoustics [13].

The development of a broad-based Hindi language recognition system using a hybrid approach combining rule-based and statistical approaches through Kaldi toolkits. The initial attempt was based on the Hindi language to automatically detect spoken sentences in Indian languages. In 1998, a hierarchical speech recognition device was reported which could recognize the Hindi phrases spoken with intervals [14] (Fig. 1).

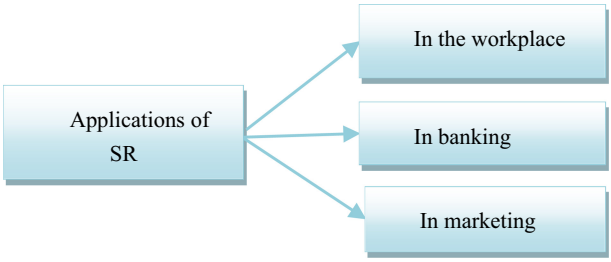

Fig. 1. Applications of speech recognition

3 Applications of Speech Recognition

3.1 In the Workplace

- Look for reports or papers on your computer;
- Make a graph or tables out of data
- Dictate the information you wish to include in a document.
- On-demand printing of documents
- Begin holding video conferences.
- Arrange meetings
- Record minutes.
- Make travel plans

3.2 In Banking

- Request information about your balance, transactions, and spending trends without opening your phone.
- Make payments
- Receive information about your transaction history

3.3 In Marketing

Voice-search has the potential to change the way marketers communicate with their customers. Marketers should be on the lookout for emerging trends in user data and behavior as people's interactions with their gadgets evolve.

4 Convolutional Neural Network

Speech recognition device that recognize words. For feature, A CNN is a DNN version often used for issues with image classification. In order to maintain spatial invariance, CNNs combine three design ideas: local receptive, mutual weights, and spatial subsampling. Consequently, CNNs are useful in three different ways relative to standard entirely linked feed-forward networks [15].

Only small local regions of an input image are linked by the local reception fields in the convolutional layers. Visual elements such as orientation borders, endpoints, and corners can be obtained from local receptive fields. Typically, nearby pixels are strongly correlated and weakly correlated with distant pixels. Thus it is structurally beneficial to stack convolutional layers to recognize images by efficiently extracting and mixing the features obtained [16].

5 Kaldi Toolkit

A scheme for the understanding of audio-visual speech using the toolkit for recognizing the Kaldi language. Kaldi was a first step in the AI conversation pipeline, which was launched at the University of Johns Hopkins in 2009 to improve technology to minimize the expense and time needed to establish speech recognition systems. This is an open-source software platform for language processing. Since then, Kaldi has been the community's de facto toolkit that allows millions of people to use expression every day. In three key steps, the standard languages to text workflow seen in the diagram below take place: extraction of features (conversion of a raw audio signal to machine learning features), acoustic modeling, and linguistic modeling. Acoustic models today at Kaldi are replaced by recurrent, convolutions-based neural networks with Gaussian Mixture (GMM) and Hidden Markov Models (HMM), to effectively simulate states, which leads to state-of-the-art outcomes [17].

6 Proposed Methodology

Propose a method to build Automatic Visual Speech Recognition System for Marathi Language (Fig. 2).

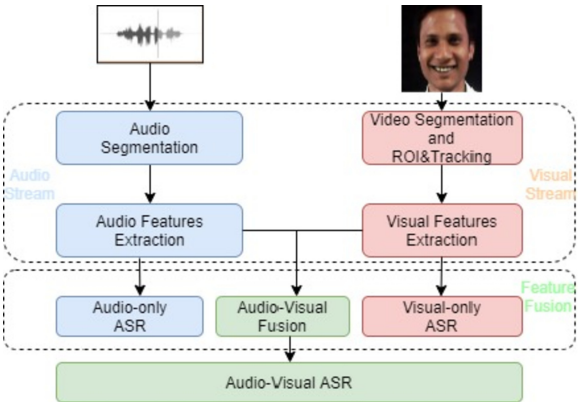

Fig. 2. Block diagram of a typical audio-video speech recognition

6.1 Combining Residual Networks with LSTMs for Lipreading

ResNets and LSTMs are the machine vision and NLP game-changers. In these books, the writers sought to offer the cumulative advantages of long and short-term memory networks in spacetime convolution, residual and bidirectional. They suggested an end-to-the-end architecture of deep learning to recognize words on visual expression, which is educated and tested on the Lip-reading In-The-Wild dataset [18].

6.2 Deep Word Embedding

The Residual and bidirectional LSTM network and the in-the-wild Lipreading database are trained. The findings reveal an average error rate of 11.92% in the 500-word language. Probabilistic Linear Discriminatory Analysis (PLDA) is used to design the embedding on words not seen during testing to make low-shot learning. The studies have shown that word-level visual perception of speech is possible even though the target words are not included in the training package [19].

6.3 Lrw-1000

A wild, naturally distributed wildlife lip-reading benchmark called LRW-1000 has been developed for researchers, containing 1,000 classes of 718,018 samples, compiled by over 2,000 single speakers. Each class corresponds to a Mandarin word syllable consisting of one or more Chinese characters [20].

Table 1. Marathi vowels

Devanagari	Transliterated	IPA
अ	a	/ə/
आ	ā	/aː/
इ	i	/i/
ई	ī	/iː/
उ	u	/u/
ऊ	ū	/uː/
ऋ	ṛ	/ru/
ए	e	/e/
ऐ	ai	/əi/
ओ	o	/o/
औ	au	/əu/
अं	aṃ	/əm/
अः	aḥ	/əɦə

6.4 Marathi Language Vowels and Consonants

The Marathi-language phoneme inventory is like that of many Indo-Aryan languages. The following is an IPA map of Marathi (Table 1).

7 Features Extraction

The videos are captured in QuickTime File Format (.MOV) with a sampling rate of 22.050 Hz using the iPhone 5. Each video has a 24-bit RGB face view of 720 × 1280 pixels. The video only captures the face of the respondent (Table 2). The participants were asked to remove eyewear or face gear during their filming procedure, and all the videos were taken in bright and dynamic environments. Audio and visual data were separated from the raw video dataset. Five visual frames taken from the video show the visual detail for each video clip. A total of 1000 video samples were taken, 1000 audio samples of 2-s length, and 5,000 visually extracted from the video samples recorded [21].

7.1 Visual Preprocessing

In order to delete data unrelated to voice and upgrade those features, visual streams must be pre-processed in order to improve speech recognition accuracy prior to application to the recognizer for testing or recognition purposes. Face identification and mouth detection and ROI extraction are the first stages in vision preprocessing. The picture that the camera has taken is an RGB picture. The picture should be changed to the gray-level picture before visual preprocessing is applied to the file. Gray-level photographs are known as monochrome images or one color. They just provide details about brightness. The standard picture consists of 8 bit/pixel (data) which allows for different levels of luminosity (0–255). The 8-bit rendering is usually because the byte that matches 8 bits of data is the tiny standard unit in the modern computing world [22].

7.2 Region of Interest (ROI) Extraction

The ROI supplies raw input data for visual function extraction and thus is affected by the correct ROI extraction in the overall output of an audiovisual automated speech recognition (AVASR) system. Due to its high deformation of the lip structure and to changes in the mouth area material due to the appearance or absence of tongue, teeth, and mouth opening and closing during the voice, the detection of ROI is made more difficult. The differences in lighting conditions and changes in speaker posture and orientation often affect ROI detection approaches. ROI extraction is often affected by the appearance or lack of a barb or mustache [23].

7.3 Visual Feature Extraction

The purpose of the extraction function is to preserve as much spoken knowledge in a relatively limited number of parameters from the original images of the speaker as possible.

Table 2. Marathi consonants

क	**ख**	**ग**	**घ**	**ङ**
ka /kə/	kha /kʰə/	ga /gə/	gha /gʱə/	ṅa (/ŋə/)
च	**छ**	**ज**	**झ**	**ञ**
ca /tɕə/	cha /tɕʰə/	ja /dʑə/	jha /dʑʱə/	ña (/ɲə/)
ट	**ठ**	**ड**	**ढ**	**ण**
ṭa /ʈə/	ṭha /ʈʰə/	ḍa /ɖə/	ḍha /ɖʱə/	ṇa /ɳə/
त	**थ**	**द**	**ध**	**न**
ta /tə/	tha /tʰə/	da /də/	dha /dʱə/	na /nə/
प	**फ**	**ब**	**भ**	**म**
pa /pə/	pha /pʰə/ or /fə/	ba /bə/	bha /bʱə/ or /və/	ma /mə/
य	**र**	**ल**	**व**	**श**
ya /jə/	ra /rə/	la /lə/	va /ʋə/	śa /ʃə/
ष	**स**	**ह**	**ळ**	**क्ष**
ṣa /ʂə/	sa /sə/	ha /ɦə/	ḷa /ɭə/	kṣa /kɕə/
ज्ञ jña /ɡɲə/				

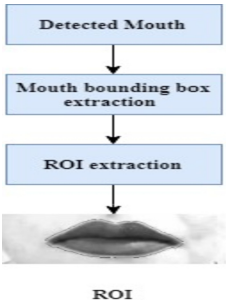

Fig. 3. ROI extraction form frame.

A variety of techniques for transformation, such as a discrete cosine transform (DCT), Discrete Wavelets Transform (DWT), and a Linear Discriminant Analytical analysis (LDA), are employed for visual feature extraction. The three detailed images are deleted and, by combining columns, the average subset image is transformed into a vector. This vector is used for the purposes of image classification as an image representation. By applying LDA, the average image in the form of ROI vectors is reduced to 30 dimensions. The DWT method in the ROI picture is shown in Fig. 3 [24].

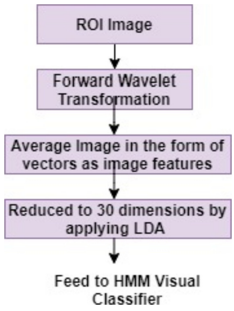

Fig. 4. DWT and HHM visual classifier

8 Conclusion

A systematic review of the current and past visual automatic speech recognition system literature (Fig. 4). What are the methodologies used up to now in this article, the original use of the Visual Speech Recognition system in this survey? It will be useful in future work to explore how the automatic visual speech recognition system can be developed for Marathi language.

References

1. Grudin, A.J.: A Moving Target—The Evolution of Human-Computer Interaction. Taylor and Francis, New York (2012)

2. Ruchika, K., Amita, D., Archana, B., Ashwani, K.: Machine learning techniques in speech generation: a review. J. Adv. Res. Dyn. Control Syst. **11**, 1095–1110 (2019). https://doi.org/10.5373/JARDCS/V11SP11/20193141
3. Hassanat, A.: Visual Speech Recognition. InTech, London (June 2011)
4. Noda, K., et al.: Audio-Visual Speech Recognition Using Deep Learning. Springer Science+Business Media, New York (2014). https://doi.org/10.1007/s10489-014-0629-7
5. Sterpu, G., Saam, C., Harte, N.: Attention-based Audio-Visual Fusion for Robust Automatic Speech Recognition. arXiv:1809.01728v3 [eess.AS]. Accessed 1 May 2019
6. Tamazin, M., Gouda, A., Khedr, M.: Enhanced automatic speech recognition system based on enhancing power-normalized cepstral coefficients. Appl. Sci. **9**, 2166 (2019)
7. Deshmukh, A.M.: Comparison of Hidden Markov Model and Recurrent Neural Network in Automatic SpeechRecognition. EJERS Eur. J. Eng. Res. Sci. **5**(8), 958–965 (2020)
8. Panda, A.K., Sahoo, A.K.: Study of Speaker Recognition Systems (2011)
9. Akkas Ali, M.D., Hossain, M., Bhuiyan, M.N.: Automatic speech recognition technique for Bangla words. Int. J. Adv. Sci. Technol. **50** (2013)
10. Gales, M., Young, S.: The application of hidden Markov models in speech recognition. Sign. Process. **1**(3), 195–304 (2007)
11. Dua, M., Saini, P., Kaur, P.: Hindi automatic speech recognition using HTK (June 2013)
12. Bansal, P., Dev, A., Jain, S.B.: Automatic speaker identification using vector quantization. Asian J. Inform. Technol. **6**(9), 938–942 (2007)
13. Mahalakshmi, P., Muruganandam, M., Sharmila, A.: Voice recognition security system using Mel-frequency cepstrum coefficients. Innovare Acad. Sci. **9**(Suppl), 3 (2016)
14. Kumar, M., Rajput, N., Verma, A.: A large-vocabulary continuous speech recognition system for Hindi. IBM J. Res. Dev. **48**(5.6), 703–715 (2004)
15. Gu, J., et al.: Recent Advances in Convolutional Neural Networks. arXiv:1512.07108v6 [cs.CV]. Accessed 19 Oct 2017
16. Chelaru, M.I., Dragoi, V.: Negative correlations in visual cortical networks. Cereb Cortex. **26**(1), 246–256 (2016)
17. https://kaldi-asr.org/doc/history.html
18. Young, T., Hazarik, D., Poria, S., Cambria, E.: Recent Trends in Deep Learning Based Natural Language Processing. arXiv:1708.02709v8 [cs.CL]. Accessed 25 Nov 2018
19. Stafylakis, T., Tzimiropoulos, G.: Zero-shot keyword spotting for visual speech recognition in-the-wild. In: 15th European Conference, Munich, Germany, Proceedings, Part IV, 8–14 Sept (2018)
20. Yang, S., et al.: LRW-1000: A Naturally-Distributed Large-Scale Benchmark for Lip Reading in the Wild. arXiv:1810.06990 [cs.CV]. Accessed 24 Apr 2019
21. Pramuk, B.: Proposal of a review valuable advices on the theory of sup- of cosmetics items. In: Proceedings of 2019 4th International Conference on Information Technology (2019)
22. Hassanat, A.B.: Visual Words for Automatic Lip-Reading. arXiv:1409.6689 [cs.CV]. Accessed 17 Sep 2017
23. Lu, Y., Li, H.: Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory. Appl. Sci. **9**(8), 1599 (2019)
24. Gautam, B.: Image Compression Using Discrete Cosine Transform and Discrete Wavelet Transform (2010)