ISSN 2278-3091
Volume 8, No.5, September - October 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse31852019.pdf
https://doi.org/10.30534/ijatcse/2019/31852019

The Performance and Classifications of Audio-Visual Speech Recognition by Using the Dynamic Visual Features Extractions

Muhammad Ismail Mohmand[1], Amiya Bhaumik[2], Muhammad Humayun[3, 4], Qayyum Shah[4]
[1]Research Scholar in the Faculty of Engineering Based Lincoln University College (LUC), Wisma Lincoln, No. 12-18, Jalan SS6/12, off Jalan Perbandaran 47301, Petaling Jaya, Selangor, Malaysia. ismail.mohmand@lincoln.edu.my
[2]Professor at the Faculty of Engineering based Lincoln University College (LUC), Wisma Lincoln, No. 12-18, Jalan SS6/12, off Jalan Perbandaran 47301, Petaling Jaya, Selangor, Malaysia. amiya@lincoln.edu.my
[3, 4]Assistant Professor at the Department of Basic Science & Islamyat University of Engineering and Technology U.E.T, Peshawar, Pakistan. humayunchemist@uetpeshawar.edu.pk
[4]Lecturer at the Department of Basic Science & Islamyat University of Engineering and Technology U.E.T, Peshawar, Pakistan. qshah08@gmail.com

ABSTRACT

Performance and classifications of the human speech recognition is bi-modular in nature and the expansion of visual data from the speaker's mouth area has been appeared to expand the presentation of the automatic speech recognition ASR frameworks. The actual performance and classifications of the audio visual speech recognitions break down quickly within the sight of even moderate commotion, however can be high quality by including visual data from the speaker mouth region. Therefore, the new methodology taken in this paper is to consolidate dynamic data caught from the speaker mouth happening during progressive casings of video got during expressed discourse. Furthermore, the audio only, visual only and audio visual recognizers were contemplated within the sight of commotion and demonstrate that the broad media recognizer has increasingly dynamic implementation.

Key words: Automatic Speech Recognition ASR, Audio-Visual Speech Recognition AVSR, Region of Interest, Hidden Markov Model.

1. INTRODUCTION

Recent investigation automatic speech recognition ASR has the principle motivation behind making human computer interaction increasingly common and succinct. Notwithstanding effective automatic speech recognition frameworks have been built up that can perform well under perfect conditions, there remains a generous test in creating arrangements that work in pragmatic circumstances where different sources or commotion are available [1]. Under such conditions, the presentation of automatic speech recognition frameworks that utilization exclusively sound data corrupt quickly, though human discourse acknowledgment, with our capacity to enhance sound with visual data, stays less seriously influenced. Various ongoing distributions have announced enhancements in discourse acknowledgment execution by joining visual data from a speaker's face or mouth area [2].

To remove reasonable visual data, explore methodologies portrayed in the writing utilize either low-level appearance highlights got from an appropriate change of pictures acquired from the speaker's mouth or face districts, or abnormal state highlights dependent on geometry, such as, length, width or roundness of mouth. In spite of the fact that the quick places of articulator's yields valuable data about verbally expressed words, these highlights neglect to catch the dynamic data present in discourse. For example, situation of tongue when expressing /I/either/d/ seems comparative, however the phonemes can be maybe better recognized by investigating the tongue's movement [3, 4].

The new methodology depicted in this paper is to consolidate data gotten from elements in the mouth region of interest that happen in progressive casings of video got during articulated discourse. The new visual features got in this work are joined with sound

features got from Mel-Frequency cepstral coefficients MFCCs and its first as well as second subordinates. Sound just, visual-just and various media recognizers have been contemplated within the sight of clamor [5, 6].

2. BACKGROUND

Appearance based systems are regularly ready to utilize an estimated ROI that necessities to bound the real mouth locale, yet in geometric based procedures a progressively precise mouth shape is required. In the geometric based methodology, the region of interest extraction and highlight computation arranges regularly turned out to be amalgamated into a solitary stage.

While the sorts of visual highlights removed fall comprehensively into three classes. In low-level or appearance based methods, the entire mouth or face locale is considered as containing discourse data. To decrease the dimensionality, an appropriate change of the speaker's mouth locale is taken trailed by principal component analysis as well as linear discriminant analysis. The geometric based procedures utilize geometric parameters of the mouth as a list of capabilities. A third method utilizes a blend of over two sorts of highlight.

In the writing, three techniques for mix have been proposed. The first is early or highlight coordination where sound and visual streams are joined at the component level. The second is late joining where the acknowledgment of the sound and visual streams are performed independently and the mix completed in the choice stage, for instance as appeared in figure 1. A third methodology is to perform halfway reconciliation in each of the early and late stages.

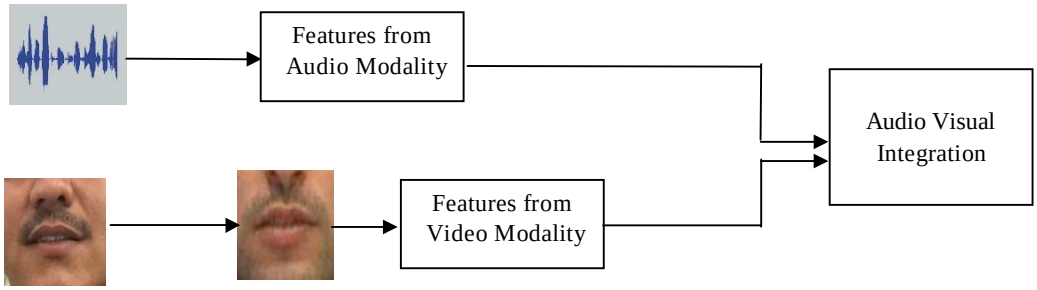

Figure 1: the block diagram of the Audio-Visual Speech Recognition AVSR system.

2.1 Audio Visual Database Techniques

A few any reasonable various media databases are accessible for discourse acknowledgment purposes, maybe halfway because of the enormous limit prerequisites for video and the noteworthy of the speaker's personality. Moreover, a few databases contain data increasingly suitable for a particular methodology; for instance databases proposed for geometric methodologies require exact localization of lip edges and corners and stamping is regularly added to the speakers' lips. As opposed to sound just ASR, no standard database is accessible for AVASR. There are just two Audio Visual Speech Analysis databases in like manner utilization, specifically the audio visual TIMIT [7, 8, 9, 10] as well as video visual Vid-TIMIT databases, the two of which contain enormous vocabularies and are reasonable for adjustment to different errands, for example, phoneme and viseme acknowledgment.

In our investigations, a subset of the Vid-TIMIT database comprising of the 30 speaker’s (15 male and 15 female) is utilized. Eight unique sentences are spoken by every speaker, containing 920 words altogether from which 22 speakers with 214 sentences are utilized for preparing and the rest of the 8 speakers with 42 sentences kept for testing purposes. Therefore, video in the database is provided at a pace of 22 outlines for every second and at a goals of 512x384. Sound is put away at 32 kHz at a profundity of the 16 bits as well [11, 12, 13, 14].

2.2 Face and Mouth Detection and Extraction Techniques

Sound highlights are separated at a pace of 100 times each second while the first video stream is 25 outlines for every second. To synchronize the sound and visual streams, the video is up-examined to 100 edges for every second utilizing straight interjection. Nearby consecutive mean quantization's transforms highlights were utilized to find the face district in the picture. The lower half of face locale is then expected to comprise the mouth area as well as a jumping box of 100x75 pixels at the focal point of these directions is removed to turn into the visual region of interest. Because of the idea of the video streams and to decrease computational cost, this procedure is just connected in the main casing of the grouping and similar directions are utilized for region of interest extraction in the rest of the edges. This methodology was effective in by far most of cases, yet once in a while the face area was either not appropriately found or the mouth locale not contained completely inside the bouncing box thus manual amendment was connected in such cases, as appeared in the Figure 2 respectively.

(i) (ii) (iii)

Figure 2.: The region of interest (ROI) extraction, (i) accurately extracted Region of Interest (ii) Missed Region of Interest (iii) Manually corrected Region of Interest.

3. FEATURE EXTRACTION TECHNIQUES

The choice of appropriate highlights assumes a basic job in the exhibition of discourse acknowledgment frameworks. In a perfect world, the highlights will hold all the important data required from the first flag identifying with discourse in a vector of little measurements. Plainly, a various media discourse acknowledgment framework necessitates that both audio and visual highlights are removed.

3.1 Audio Feature Extraction

In audio feature extraction the standard MFCCs are used as well as the Cambridge University Hidden Markov Model HMM Toolkit is utilized to remove 16 Mel-frequency cepstral coefficients alongside its first and second subordinates [15, 16, 17].

3.2 Visual Feature Extraction Techniques

The new dynamic based methodology utilized here for visual element extraction considers the elements of the mouth district during discourse that are not caught by the appearance-based and geometric based component strategies announced in writing. Movement vectors are determined between progressive edges utilizing the square coordinating calculation depicted in [18, 19]. The region of interest separated in segment 4 is resized to 88x70 pixels so as to permit an essential number of large scale squares of size 6x6 to be produced. As the required discourse data is for the most part present in vertical developments, the 90 measurement movement vector is acquired by reshaping just the vertical segments of the got movement vectors. The principle component analysis is then connected to lessen the quantity of measurements to 30 and this vector is utilized related to the 6 sound highlights to give an early coordination approach [20, 21, 22, 23].

4. RESULTS AND CONCLUSION

In these analyses, the Hidden Markov Model and toolkit of the Cambridge University has been embraced for preparing and testing purposes and no utilization is made of word reference data or a language model during the acknowledgment procedure.

Three separate speech acknowledgment frameworks were prepared for sound just, visual just and broad media discourse acknowledgment. Investigations were performed with a scope of sound commotion levels. As can be seen from Table 1, the sound just recognizer beats both visual-just and broad media recognizers when no clamor is available, yet its relative execution decays as extra commotion is presented. True to form, the presentation of the visual-just framework is free of sound clamor and the various media acknowledgment framework is more vigorous to commotion than the sound just strategy.

Table 1: The comparative performance of audio visual speech recognition AVSR system

signal to noise ratio	Audio only	Visual only	Audio visual
clean speech	33.98	27.35	28.68
20 db	33.96	27.35	28.68
15 db	33.69	27.35	28.16
05 db	33.15	27.35	28.16
0 db	23.13	27.35	28.89
-05 db	23.05	27.35	28.55

REFERENCES

1. Miyajima, C., Tokuda, K., and Kitamura, T. (2000). Audio-visual speech recognition using MCE-based HMMs and model-dependent stream weights. *Proc. International Conference on Spoken Language Processing*, Beijing, China, vol. II, pp. 1023-1026.
2. Movellan, J.R. and Chadderdon, G. (1996). Channel separability in the audio visual integration of speech: A Bayesian approach. In Stork, D.G. and Hennecke, M.E. (Eds.), *Speech reading by Humans and Machines*. Berlin, Germany: Springer, pp. 473-487. https://doi.org/10.1007/978-3-662-13015-5_36
3. Nadas, A., and Picheny, M. (1989). Speech recognition using noise adaptive prototypes. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 37*:1495-1503. https://doi.org/10.1109/29.35387
4. Nakamura, S., Ito, H., and Shikano, K. (2000). Stream weight optimization of speech and lip image sequence for audiovisual speech recognition. *Proc. International Conference on Spoken Language Processing*, Beijing, China, vol. III, pp. 20-23.
5. Nakamura, S. (2001). Fusion of audio-visual information for integrated speech processing. In Bigun, J. and Smeraldi, F. (Eds.), *Audio-and Video-Based Biometric Person Authentication*. Berlin, Germany: Springer-Verlag, pp.127-143. https://doi.org/10.1007/3-540-45344-X_20
6. C. Gaida, P. Lange, R. Petrick,. Malatawy and D. Suendermann Oeft, "Comparing open-source speech recognition toolkits," Tech. Rep., DHBW Stuttgart, 2014.
7. Y. V. Varshney, Z. A. Abbasi, M, and O. farooq "Frequency Selection Based Separation of Speech Signals with Reduced Computational Time Using Sparse NMF," Archives of Acoustics, Vol. 42, No. 2, 2017, pp. 287-295. https://doi.org/10.1515/aoa-2017-0031
8. D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," Nature, 2nd ed. vol. 401, no. 6755, 1999, pp. 788- 91. https://doi.org/10.1038/44565
9. P. O. Hoyer, "Non-negative Matrix Factorization with Sparseness constraints," Journal of machine learning research, vol. 5, 2004, pp. 1457-1469.
10. P. Upadhyaya, O. Farooq, and M. R. Abidi. "Continuous Hindi Speech Recognition Model Based on Kaldi Automatic Speech Recognition Toolkit" in IEEE conference WiSPNET, 2017, pp. 812-815. https://doi.org/10.1109/WiSPNET.2017.8299868
11. K. Kumar, C. Kim, and R.M. Stern "Delta-Spectral Cepstral Coefficients MFFCs for robust speech recognition," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),2011, pp. 4784-4787. https://doi.org/10.1109/ICASSP.2011.5947425
12. J. Luettin, N. A. Thacker, and S. W. Beet. Speech reading using shape and intensity information. In Proceedings of the 4th International Conference on Spoken Language

Processing (ICSLP'96), volume 1, pages 58-61, 1996.

13. J. Luettin, N. A., and S. W. Beet. Statistical lip modelling for visual speech recognition. In G. Ramponi, G. L. Sicuranza, S. Carrato, and S. Marsi, editors, Signal Processing VIII Theories and Applications, volume I, pages 137-140, Trieste, Sept. 1999.
14. J. Luettinga, N. A. Thacker, and S. W. Beet. Visual speech recognition using active shape models and hidden markov models. In Proc. International Conference on Acoustics, Speech recognition and Signal Processing, volume 2, pages 817-820, Atlanta, GA, May 1996. IEEE.
15. D. W. Massaro and D. G. Stork. Speech recognition and sensory integration. American Scientist, 86, May 1998. https://doi.org/10.1511/1998.3.236
16. I. Matthews, J. A. Bangham, R. Harvey, and S. Cox. A comparison of active shape model and scale decomposition based features for visual speech recognition. In Proc. European Conference on Computer Vision, pages 514-528, June 1998. https://doi.org/10.1007/BFb0054762
17. W. Hürst, and P. Duchnowski. Adaptive bimodal sensor fusion for automatic speech reading. In Proc. International Conference on Acoustics, Speech and Signal Processing, volume 2, pages 833-836, Atlanta, GA, May 1996. IEEE.
18. W. Kim and J. H. L. Hansen, "Feature Compensation Employing Variation Model Composition for Robust Speech Recognition in In-Vehicle Environment," in Digital Signal Processing for In-Vehicle Systems and Safety, J. H. L. Hansen, P. Boyraz, K. Takeda, and H. Abut, Eds. Springer United States US, 2012, pp. 175-185. https://doi.org/10.1007/978-1-4419-9607-7_11
19. X. Zhang, C. C. Broun, and M. a. Clements, "Automatic Speech reading with Applications to Human Computer intelligent Interfaces techniques," EURASIP J. Advances in Signal Process and recognition., vol. 2002, no. 11, pp. 1228-1247, 2002. https://doi.org/10.1155/S1110865702206137
20. H. Sahbi, and G. Vito, "Designing relevant features for visual speech recognition," in 2013 IEEE International Conference on Acoustics automatic Speech reading and Signal Processing, 2013, pp. 2422-2426.
21. C. Neti, J. Luettin, and I. Matthews, "Audio visual automatic speech recognition AVASR: An overview," Issues in Visual and Audio Visual Speech Processing techniques, Massachusetts Institute of Technology, MIT Press, 2009.
22. Potamianos, J. Matthews, H. Glotin, and D. Vergyri, "Large vocabulary audio visual speech recognition AVSR: A summary of the Johns Hopkins Summer 2000 Workshop" in Processing Works. Multimedia Signal Processing, 2009, pp. 618-623.
23. Mohammed Alameri, Osama Isaac, and Amiya Bhaumik. Factors Influencing User Satisfaction in UAE by using Internet. Published in International Journal on Emerging Technologies, 2019, volume 10, Issue -1a. 8-15 Pages.