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/ijatcse27852019.pdf
https://doi.org/10.30534/ijatcse/2019/27852019

An Assessment of the Visual Features Extractions for the Audio-Visual Speech Recognition

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

Utilization of the visual data from the speakers mouth region has appeared to develop presentation of the Automatic Speech-Recognition ASR frameworks. This is the particularly valuable in nearness of the clamor, which uniform in the moderate structure seriously debases discourse acknowledgment execution of frameworks utilizing just sound data. Different arrangements of highlights separated from speakers mouth area have been utilized to improve the showing of an ASR framework. In such testing situations and have met various triumphs, and to the best of creators information, the impact of utilizing these methods on the acknowledgment execution based on the phonemes have not been examined at this point. This paper presents examination of phoneme acknowledgement execution utilising visual highlights removed from mouth area of-enthusiasm utilising discrete cosine transform and discrete wavelet transform. Therefore, new discrete cosine transform and discrete wavelet transform feature have likewise been extricated and contrasted and the recently utilized one. These highlights were utilized alongside sound highlights dependent on the Mel-Frequency Cepstral Coefficients MFCCs. This recent research will help in the choosing appropriate feature for various application as well as distinguish the restrictions of these techniques in the acknowledgment of the individual-phonemes.

Key words: Audio Visual Speech Recognition (AVSR), Motion Vector, Hidden Markov Model.

1. INTRODUCTION

The performance of ASR has pulled in a ton of the enquiries during previous couple of decades to make the man-machines connection progressively normal. The automatic speech recognition abilities are accounted for to have progressed to close to human's degree of acknowledgment, this is commonly just feasible under perfect conditions and the exhibition falls apart altogether within the sight of sound clamor [1].

To expand the ability of current automatic speech-recognition frameworks and make it vigorous to commotion, visual discourse is a characteristic up-and-comer. It has for some time been realized that the visual data from the speakers mouth locale improve discourse acknowledgment by the people in nearness of clamor [2]. Anyway the utilization of sound as well as visual modality for automatic speech recognition known as Audio-Visual Speech Recognition AVSR system. In this work geometric parameters were removed from highly contrasting pictures of the mouth area of the speaker. This was trailed by various examinations concerning other enlightening highlights from speaker's mouth locale for AVSR [3].

AVSR framework comprise of two channels of data for example sound and video-channels, and the fuse of the visual data is related with the new errands which can be sub-divided into face subsequent as well as mouth locale of intrigue region of interests ROIs extractions, visual feature extractions and broad media combination [4].

A few calculations have-been planned for the face subsequent and mouth extractions, including shading based division. In any case, these strategies as a rule accomplish sub ideally in shifting lighting condition. The AVSR task-rather requires an increasingly exact gauge of lip parameters for getting an adequate exhibition. Hence, the speaker's mouths are shaded either some stamping is set on it, to support the exact subsequent as well as the estimation of the lip shape [5]. Therefore for incorporation of sound and visual-floods of data; there are three primary strategies specifically, initial combination achieved on the highlights level, late reconciliation did at choice levels and techniques that fall among these two limits.

Brief performance of the AVSR frameworks is incredibly reliant on the extractions of the visual feature that hold however much data as could reasonably be expected about the first pictures that is important to discourse acknowledgment. Geometric based visual highlights, for example, mouth opening and shutting, mouth tallness, width and region and so on have been utilized. The greater part of the systems utilized in this sort of highlights uses hued lips or other stamping [6], anyway this methodology is a long way from this present reality circumstances. Procedures for programmed lip form extraction have been proposed by a few creators, however to-date these have met with restricted achievement [7] and an elective methodology is to receive appearance based features extractions strategies, which can be apply reasonable change of the mouth region of interest pursued by dimensionality decrease systems, for example linear discriminant analysis as well as principle component-analysis [8]. This article writes about a correlation of visual highlights separated utilizing an appearance methodology. The new features are separated from groups of three-dimensional frequencies in the discrete cosine transform as well as discrete wavelet-transform areas and their presentation contrasted and existing methodology of utilizing entire scope of the spatial-frequencies for the highlights extraction.

3. DATABASE TECHNIQUES

Dissimilar to sound just discourse acknowledgment where standard databases are proliferate just few databases are accessible for broad media discourse acknowledgment. This is because of various factors, for example, the overall intricacy of securing and preparing broad media information, the more noteworthy stockpiling necessity for video stream and that AVSR research is for the most part completed by individual specialists or little gatherings of scientists. The databases that are accessible regularly experience the ill effects of poor video quality, predetermined number of speakers, as are not appropriate for consistent discourse acknowledgment tests. Apparently there are right now two databases appropriate for enormous vocabulary AVSR undertakings in particular audio visual TIMIT as well as VidTIMIT databases [9]. VidTIMIT database covers 40 speaker's (22 guys and 18 females) as well as subset of this database having 30 speaker's (15 guys and 15 female's speaker's) was utilized in work depicted in this article. Every speaker expresses eight distinct sentences before a camera fixated on substance of speaker, and the sentences in database are on the whole instances of persistent discourse booked from the standard VidTIMIT database as well as comprise an aggregate of the 210 expressions and the terms of 920 words, and the sound is recorded at the test rate of 64 KHz and 32 bits profundity; video is recorded at the rate of 24 outlines for each second [10].

2. REGION OF INTEREST ROI EXTRACTIONS

The features extractions are removed from the sound-steam multiple times of the each second. The video casings were up-examined to the rate of 50 casing for every second utilizing direct insertion. Limited Successive Means Quantization's Transform SMQT [11] highlights were utilized to discover the face-region of the images. The face-region is the recognized in the main edge of articulation techniques and the instructions of lower half part are resolved. Therefore, a 100x70 region focused on these directions are separated as mouth region of interest and these equivalent directions are utilized for region of interest extractions in the rest of casings of the expression [12, 13, 14]. This methodology was found to function admirably when all is said in done and furthermore lessens the time required for

extricating mouth area in each picture independently, and one such mouth locales in this method separated is appeared in the Figure 1 (i). In modest quantity of situations where the mouth region was not precisely situated by this procedure, physical remedy is finished. Figure 1 (ii) demonstrates one missed face, while in Figure 1 (iii) demonstrates the equivalent redressed physically.

4. VISUAL-FEATURES EXTRACTION TECHNIQUES

Extraction of the visual-features that comprise an excellent data reasonable for discourse acknowledgment object is a basic stage in the AVSR system. The modern methodologies have been embraced for visual element extractions, can be assembled into classes of the appearance based, geometric based as well as cross based techniques. In this work the appearance based feature have been utilized in all directions.

4.1 Appearance Based Features Techniques

The appearance based features extractions methodologies are considered for whole mouth region of speaker to be enlightening for the discourse acknowledgment. Recurrence data is frequently significant in sign investigation and reasonable changes of the mouth region of interest are regularly taken to concentrate such data. The most generally utilized changes in picture pressure writing are the DCT as well as DWT. Therefore the DWT is just the genuine segment of Fourier change with the sign investigation being-performed at a uniform goals. While the wavelet examination perform examination at a scope of the goals for both in time as well as scale and is along these lines are known as multi goals investigation.

(i) (ii) (iii)

Figure 1: The region of interest extractions, (i) Accurately extracted of region of interest (ii) Missed region of interest (iii) Manually corrected of region of interest extractions.

4.2 Linear Discriminant Analysis

Various systems are utilized for order of information. Two ordinarily utilized information investigation procedures are principle component analysis as well as linear discriminant analysis. The principle component analysis changes information arranged by diminishing difference to such an extent that limit of fluctuation lies about the main pivot then the subsequent hub, etc. This kind of strategy is progressively valuable for speaking to information in a minimal arrangement of measurement. Be that as it may, for information arrangement, where the prime reason for existing is to segregate between various classis, this strategy isn't ideal. The linear discriminant analysis then again changes information as to boost between class difference and limits the inside class fluctuation. On the off chance that inside class disperse lattice is meant by S_w and between class dissipate framework by S_b then the change grid W is with the end goal that is augmented,

$$J(W) = \frac{WSbW}{WSwW}$$

The ideal W comprises of the eigen-vectors comparing to k is the biggest Eigen esteems, where k is belongs to an ideal dimensionality reduction of changed space.

5. EXPERIMENTS

As talked about before in the region of interest, the face area is found utilizing progressive mean quantization change and mouth-region of size 100x70 are extricated about the focal point of it's lower half and mouth-region subsequently separated is re-sized to measure 60x60 different four locales. The two dimensional discrete cosine transform and discrete wavelet transform are taken and isolated into four recurrence regions as appeared in the Figure 2. Therefore, in the primarily arrangement of the examinations 100 most elevated vitality-coefficients

are extricated from four areas both for discrete cosine transform as well as discrete wavelet transform coefficients, trailed by linear discriminants analysis to get our video-perception vector of the 40 measurements, while in the second trial 20x20 region are reformed to a vector of the 400 measurements and pursued by the Linear Discriminant Analysis to achieve the last arrangement of the 40 measurements. The class marks for the Linear Discriminant Analysis step are given by bootstrapping on sound just Hidden Markov Model HMM grew before utilizing constrained arrangement.

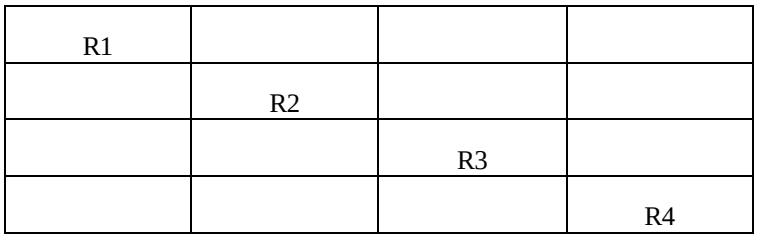

Figure 2: Regions for the features selection techniques

6. HIDDEN MARKOV MODEL MODELING TECHNIQUES

Perception vector gotten by our feature extractions techniques talked about above structure a 20 dimensional static element. Therefore, the delta and delta-delta features are added for a dynamic visual component vector of the measurement 60. In our comprehensive media tests 16 Mel-Frequency Cepstral Coefficients and their first as well as second subordinates are extricated and added to ninety measurement of the dynamic visual component vector accordingly offering ascend to an early combination methodology technique, and three main states of the Hidden Markov Model HMM as appeared in the Figure 3, is created for phonemes set of 50 alongside their setting subordinate tri-telephone model by utilizing Cambridge university toolkit as well.

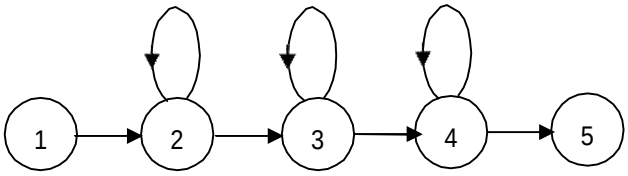

Figure 3: Hidden Markov Model with tree emitting states techniques.

7. RESULTS AND DISCUSSION

In spite of the fact that the examinations have-been performed on broad audio and video only assignments yet here the outcomes on video only investigations are accounted. An explanation behind the language-model is just founded on a phoneme premise and joining language-model maybe influence the outcomes which won't be actually a proportion of execution of it's video segments. At this time we look at the outcomes over various elements. We contrast discrete cosine transforms based coefficients and discrete wavelet transforms based partners. As is obvious from Figure 4 the discrete cosine transform based highlights outflank their partner in discrete wavelet transform based set and again an examination of utilizing the vitality based feature with utilizing the entire list of capabilities is performed. Clearly utilizing the entire arrangement of coefficients for highlights extraction utilizing linear discriminant analysis when all is said in done gives preferred outcomes over utilizing high vitality coefficients. A conceivable contention in help of utilizing vitality based coefficient might be its lower-measurements however as preparation is done disconnected and impacts of utilizing entire highlights set will have negligible impact on acknowledgment time. All over again the examination of various recurrence districts gives that middle of the road frequencies are more enlightening for discourse acknowledgment than low vitality highlights.

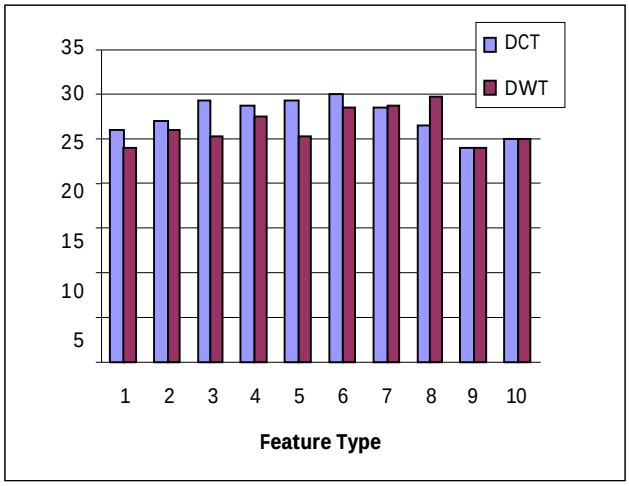

Figure 4: Results for the discrete cosine transform as well as discrete wavelet transform based features extractions techniques

1. R1 Energy: 2. R1 whole, 3. R2 energy, 4. R2 whole, 5. R3 energy, 6. R3 whole, 7. R4 energy, 8. R4 whole, 9. Full-energy, 10. Full-whole.

8. CONCLUSION

This research examines the exhibition of the various visual features for programmed discourse acknowledgment. Because of utilization of various databases by broad media enthusiastic discourse acknowledgment network as well as non-presence of the standard-face and mouth extractions methods our outcomes cannot be contrasted legitimately and some other research regarding this matter. In this work, video-TIMIT databases are utilized for the persistent discourse acknowledgment task, which comprise of generally enormous quantity of the subjects. Then we contrast our outcomes and procedures announced in the various media passionate discourse acknowledgment writing, on utilizing our-own trial setup. The proposed locales-based features are novel and are contrasted and highlights separated from entire arrangement of the constants are changed space demonstrated in the results as full-picture. Outcomes are accounted for now on the visual just speech acknowledgment with no utilization of the language-model. This gives an immediate correlation of the visual highlights with non-other factors are included. The outcomes is demonstrate that the discrete cosine transform based-features as a rule gives better execution thought about than wavelet change based highlights. Again we see that low recurrence coefficients in spite of the fact that gives best execution for picture reclamation yet with the end goal of discourse acknowledgment halfway frequencies gives improved execution and this maybe the cause that middle of the way levels features comprises added data about the lip minute which are necessary for the discourse acknowledgment. Utilizing the entire list of capabilities for preparing rather than high vitality coefficients additionally add to execution of discourse acknowledgment framework. Utilization of entire list of capabilities in spite of the fact that expansion of dimensionality reduction and expands preparation in time, however it's impact on the speech acknowledgment is insignificant as preparation procedure is performed disconnected.

REFERENCES

1. G. Young, S. Everman, N. Gals, T. Haim, H. Kershawa, Y. Lu, H. More, J. Odel, D. Olason, D. Poveye, V. Valtchv, G. Woodlands, "The Hidden Markov Tool Kit HTK Book, (for version 3.5)". *The*

Cambridge University Engineering Department, 2008.

2. A. Leees, T. Kawaharaa, K. Shikanu. "Julius an open source real time large vocabulary recognition engine system". *EURO-SPEECH*, pp. 1692-1695, 2009.
3. N. Walker, P. Lameres, P. Kwoka, B. Raja, G. Singh, E. Gouge, P. Wolfs, J. Woolfell. "Sphinx-4: A Flexible Open Source Framework for Speech-Recognition". The *Sun Microsystems Inc. Technical Report, SML1, TR2004 0811*, 2004.
4. H. Reach, C. Golan, G. Rheingold, B. Hofmeister, J. Loaf, R. Schulte, H. Neye. "The RWTH Aachen University of Science Open Source Speech-Recognition System". *INTERSPEECH*, pp. 2112-2115, 2009.
5. G. Running, Z. Peie, G. Qiny, Z. Zipping, W. Hai, W. Xining. "CASA Based Speech Separation for Robust Speech-Recognition". *Proceedings of the Ninth International Conference on the Spoken Language Processing, (ICSLP)*. pp. 2-6, 2008.
6. Y. U. Vashney, Z. A. Abbasid, M. R. Abide, O. Farooq, F. Upadhyay. "SNMF Based Speech Demising with the Wavelet Decomposed Signal-Selection". *IEEE conference WISPNET*, pp. 2644-2648, 2018. https://doi.org/10.1109/WiSPNET.2017.8300234
7. U. Q. Wange, D. Wange. "A Joint Training Framework for the Robust Automatic Speech-Recognition ASR ". *IEEE and ACM on Transactions on Audio-Speech and Language Processing*, vol. 25, no. 5, pp. 799-809, April 2018.
8. F. Upadyaya, O. Farooq, M. Abide, P. Varshny. "Comparative Study of the Visual Feature for Bimodal Hindi Speech-Recognition". *Archives of the Acoustics*, vol. 41, no. 5, pp. 608-618, 2016.
9. F. Mowlem, R. Saied, M. G. Christensena, Z. H. Tang, T. Kinnuen, P. Frantic, S. H. Jensen. "A Joint Approach for Single channel Speaker Identification and Speech-Separation". *IEEE Transactions on the Audio Speech and Language Processing techniques*, vol. 22, no. 8, pp. 2588-2603, 2013.
10. N. Khadeian, M. Homayunpour. "Monaural Multi Talker Speech Recognition by using Factorial Speech Processing Models". *CoRR*, 2017. https://doi.org/10.1016/j.specom.2018.01.007
11. F. Young, G. Everman, M. Gale, T. Haim, D. Kersha, X. Li, G. More, J. Odel, D. Olsson, D. Pokey, V. Vetches, P. Woodland. "The Hidden Markov Tool Kit HTK Book (for version 3.6)", *Cambridge University Engineering Department*, 2011.
12. U. Krcadinac, Pasquir. "Synesketch an Open Source Library for Sentence Based Emotion Speech-Recognition". An Affective Computing IEEE Transactions on, vol.3, no.3, pp.313, 328, July to Sept, 2015.
13. B. Hyvainen and E. Ojai. “Independent Component Analysis: Algorithms and Applications”. The neural networks, vol. 14, no. 4-6, pp. 412–432, 2009.
14. 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.