ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 5, May 2013

Enhancement of ASR using viseme clue

Priyanka Sharma[1], Parul Dihulia[2], Vikas Gupta[3]

Research Scholar, EC Department, TIT, Bhopal, Madhya Pradesh, India [1]

Assistant Professor, EC Department, TIT, Bhopal, Madhya Pradesh, India [2]

Head of Department, EC Department, TIT, Bhopal, Madhya Pradesh, India [3]

Abstract: Automatic Speech Recognition (ASR) is an essential component in many Human-Computer Interaction systems. A variety of applications exits in the field of ASR and shown a good improvement and reached high performance levels but only for condition-controlled environments. Hence a robustness technique has to be build for the ASR which can improve the performance of ASR in the adverse environment.

Keywords: ASR, MFCC, 2-DWT, AVASR, Visual clue

I. INTRODUCTION

Today the most commonly natural communication tools used by humans are their voice. It can be consider in such sense that thee lot of research have been put into the field of Automatic speech recognition system (ASR). The goal of Automatic speech recognition is to transcribe a recorded speech utterance into its corresponding sequence of words [1].

Another application that exit is speaker recognition, where the goal is to determine either the claimed identity of the speaker (verification) or who is speaking (identification). Though the lot of contribution [1- 2] has been made in this area still an enormous amount of research has been devoted to speech processing, there appears to be some form of local optimum in terms of the fundamental tools used to approach these problems.

It has been [3] reported that speech recognition by machines is by far not as robust as recognition by humans and human listeners apparently face little or more difficulty in the presence of background noise.

Hence the machine performance degrades quite rapidly if the speech signal is degraded by acoustic environmental noise, reverberation, competing speakers or any other kind of distortion. Further development is necessary for better accuracy in real conditions, where environmental or other kind of noises exist. Audio signal features need to be enhanced with additional sources of complementary information to overcome problems due to large amounts of acoustic noise. This lack of robustness has been identified as one if not the major impediment to the ubiquitous use of automatic speech recognition (ASR) technology. Here in this paper we have proposed a scheme in which visual clue along with the audio feature are integrated and passed through the system and hence Audio-Visual Speech recognition system (AVASR) is developed.

II. LITERATURE REVIEW

The research [4-5] carried out in the field of AVASR reported the increased in performance of ASR using the visual clues which make our speech more easily readable in the adverse environment resulting in audio-visual speech recognition.

Visual information [5-8] extracted from speaker's mouth region seems to be promising and appropriate for giving audio-only recognition a boost. Region of interest i.e. lip region is selected from the facial expression using traditional image processing methods [7] and combining with audio feature [8].

Color-based detection [3- 8] strategies are used to detect and track the mouth region, which is considered as the Region of Interest (ROI) through sequential time frames. Subsequently, Discrete Cosine Transform (DCT) is used along with the Discrete Wavelet Transform for extracting the visual feature. Furthermore, Audio and Visual stream fusion appears to be even more challenging and crucial for designing an efficient AV Recognizer.

In this paper we have investigate that how the viseme clue can be helpful in increasing the overall performance of the Audio Speech recognition. So in this paper we have selected some of the viseme classes and then the performance of audio-only recognition in noisy environment and audio-visual recognition is carried out.

On the other hand various speech enhancement techniques are been proposed [9] based on the criteria used or application of the enhancement system. The speech signal can be acquired from single or multiple channel sensors. Additive noise can make speech enhancement particularly difficult [2]. Non-stationary of the noise process can further complicate the enhancement effort [3-4].

On the other hand suppression of noise using periodicity of speech methods exploit the quasi-periodic nature of voiced speech [1]. Voiced speech is periodic in nature characterized by a fundamental frequency, which varies from person to person. This technique however, depends heavily on the accurate estimation of the pitch period (inverse of the pitch) of the speaker's voice.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 5, May 2013

Another method based on the adaptive comb filter [10] in which a series of notch filters are used so as to filter out any spectral content between the fundamental frequency and its harmonics. Another method is the single channel adaptive noise cancellation technique [11].

Model-based [12] speech enhancements are also called statistical model based methods; these methods are usually used when there is no knowledge of the statistical properties of the speech or noise signal. Spectral subtraction [13] is a well-known noise reduction method based on the STSA estimation technique. The basic power spectral subtraction technique, as proposed by Boll [14], is popular due to its simple underlying concept and its effectiveness in enhancing speech degraded by additive noise. But there was some drawback with the spectral subtraction methods such as

Residual noise (musical noise) in which there is some significant variations between the estimated noise spectrum and the actual noise content present in the instantaneous speech spectrum. The subtraction of these quantities results in the presence of isolated residual noise levels of large variance. Several residual noise reduction algorithms [1-3] have been proposed to combat this problem. However, due to the limitations of the single-channel enhancement methods, it is not possible to remove this noise completely, without compromising the quality of the enhanced speech. Hence there is a trade-off between the amount of noise reduction and speech distortion due to the underlying processing.

Roughening of speech due to the noisy phase in this signal is not enhanced before being combined with the modified spectrum to regenerate the enhanced time signal. This is due to the fact that the presence of noise in the phase information does not contribute immensely to the degradation of the speech quality.

Apart from using such techniques it has been found out that the environmental robustness plays an important role in determining the success of an ASR [3, 8]. Even the performance of ASR system degrades in 'clean' acoustic environment, when the training and the test background conditions are different [3].

Hence a new technique was proposed [15] in which the audio feature is integrated with video feature. Petajan [15] first proposed the use of visual features along with audio to develop Audio-Visual ASR. Since the visual information is not affected by the presence of noise or acoustic background conditions, since then the Audio Visual Automatic Speech Recognition (AVASR) has become the major area of research interest for the researcher.

III. VISUAL SPEECH RECOGNITION

Most of the work done on VSR came through the development of AVSR systems, as the visual signal completes the audio signal, and therefore enhances the performance of these systems. Little work has been done using the visual only signal. Most of the proposed lip reading solutions consist of two major steps, feature extraction, and Visual speech feature recognition. The feature [6, 8] used in the case of visual are based on either of the technique based on Geometric features-based approaches, Appearance-based approaches, Image-transformed-based approaches and Hybrid approaches.

Classifier	A	B	C
Linear	57.77	59.32	77.96
Mahanloabis	54.23	50.84	72.03
Quad	55.93	50.00	72.88

Table 1.1 Recognition rate

A geometric features-based approach includes the first work on VSR done by Petajan in 1984, who designed a lip reading system to aid his speech recognition system. His method was based on using geometric features such as the mouth's height, width, area and perimeter [15]. Appearance-based lip reading system [16], employing a novel approach for extracting and classifying visual features termed as "Hyper Column Model" (HCM).

The image transformed [17] work was designed to be posing invariant. Their audio-visual automatic speech recognition was designed to recognize speech regardless of the pose of the head, the method starting with face detection and head pose estimation.

Hybrid feature [8] was proposed in which visual features obtained from DCT and active appearance model (AAM) were projected onto a 41 dimensional feature space using the LDA. Linear interpolation was used to align visual features to audio features.

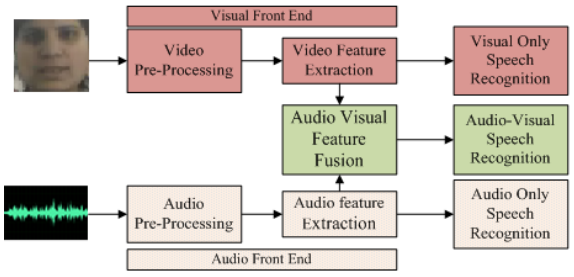

Figure 1.1 Basic procedures for AVASR

IV. EXPERIMENTAL APPROACH

The figure 1.1 shows the basic procedure for Audio-Visual Speech Recognition. It can be observe from the figure that the overall video is break into two part i.e. audio and frame (image). Therefore we have classified it into two end i.e. audio front end and visual front end.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 5, May 2013

In Audio front end the pre-processing of the signal is done to remove the redundancy if any present and thereafter the audio feature are extracted. Here in our case audio feature was extracted using the MFCC and then audio only recognition is performed.

Same procedure is repeated for the video front end in which the frame are extracted and the pre-processing is done in the same way as done for audio, but the pre-processing here is different. Here since our region of interest is only the lip region therefore only the face is extracted from the frame and finally the lip are extracted as shown in figure 1.2 and finally the DWT and DCT is performed on the ROI for taking the video feature.

After that the audio and video feature are integrated to give the audio only recognition. Before the recognition the feature are passed through the classifier where the training and testing of the feature are done to perform final recognition. Here in our case we have used the linear discernment analyser as the classifier. In our case we have selected the three viseme classes and the corresponding phoneme are extracted.

V. CONCLUSION

Using the test data as input experiments are performed on the MATLAB. Result of the program is obtained in the form of Confusion Matrices. Recognition accuracy of individual phonemes of the three viseme class is obtained by the Phoneme recognizer followed by the Viseme recognizer. Viseme recognition is carried out basically for the three viseme classes' viseme i.e. class 1, 2 and 3. The approach for the experiment was to find the audio only recognition, video only recognition and audio video only recognition.

Finally the percentage recognition for Audio only recognition (A), Video Only recognition (B), Audio Video using 49 i.e.(13 MFCC plus 36 DCT) feature (C). Table 1.1

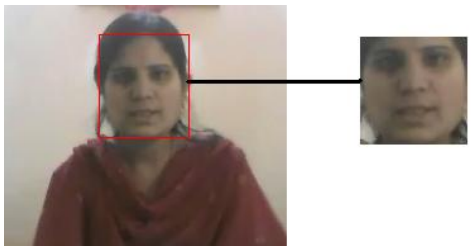

Figure 1.2 Video processing.

shows the Audio only recognition (A), Video only recognition (B), Audio Video using 49 i.e. (13 MFCC plus 36 DCT) feature(C). The experiments were carried out in three phases, in the first phase audio only recognition was carried out using 13 MFCC based features. In the second phase, video only features are extracted for recognition. And finally in third phase DCT features was integrated with 13 MFCC feature for viseme class recognition.

Table 1.1 shows the percentage recognition of the three phases of the experimental setup. According to figure, recognition rate is increased by 20.19%, 17.81% and 16.95% for Linear, Mahalanobis and Quadratic classifier respectively for the optimum value. From the above result it is concluded that the AVASR has an improvement over the audio only recognition and also over visual only recognition for clean speech

VI. REFERENCES

1. L.R Rabinder, B.Juang ,Fundamentals of Speech Recognition,183-185,Prentice Hall Inc, London ,1993.
2. C. Neti ,N. Rajput ,A. Verma ",A large –vocabulary continous speech recognition system for Hindi ," IBM Journal of Research and Development, Volume-48, Num5/6,2004.
3. Potamianos, G., Neti, C., Luettin, J., & Matthews, I. (2004). Audio-visual automatic speech recognition: An overview. In G. Bailly, E. Vatikiotis-Bateson & P. Perrier (Eds.), Issues in Visual and Audio-Visual Speech Processing: MIT Press.
4. Rowan Seymour, Darryl Stewart and Jiming, 2008. Comparison of image transform-based features for visual speech recognition in clean and corrupted videos. Hindawi Publishing Corporation, EURASIP Journal on Image and Video Processing. Volume 2008.
5. Visser, M., Poel, M. and Nijholt, A.: Classifying Visemes for Automatic Lipreading, Proc. 2nd International Workshop on Text, Speech and dialogue (TSD' 99), LNAI 1692, pp.349–352 (1999).
6. E.Bozkurt, C. Eroglu Erdem, E.Erzin, T.Erdem and M.Ozken, "Comparison of Phoneme and Viseme Based Acoustic Units for Speech Driven Realistic Lip Animation Animation", 3DTV conference, 7-9 May 2007.
7. Werda, S., Mahdi, W., & Ben-Hamadou, A. (2007). Lip Localization and Viseme Classification for Visual Speech Recognition, International Journal of Computing & Information Sciences, Vol.5, No.1.
8. Neti, C., Potamianos, G. & Luettin, J. (2000). Audio-visual speech recognition, Final Workshop 2000 Report, Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD.
9. J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," Proc. IEEE, vol. 67, No. 12, pp. 221-239, Dec. 1979.
10. C. He and G. Zweig, "Adaptive two-band spectral subtraction with multi-window spectral estimation," ICASSP, vol.2, pp. 793-796, 1999.
11. Y. Hu, M. Bhatnagar and P. Loizou, "A cross-correlation technique for enhancing speech corrupted with correlated noise," ICASSP, vol. 1, pp. 673-676, 2001.
12. Y. Ephraim, "Statistical-model-based speech enhancement systems," Proc. IEEE, vol.80, No.10, pp. 1526-1555, Oct.1992.
13. J. Deller Jr., J. Hansen and J. Proakis, "Discrete-Time Processing of Speech Signals",NY: IEEE Press, 2000.
14. S. F. Boll, Suppression of acoustic noise in speech using spectral substraction, IEEE Trans. Acoustics, Speech and Signal Processing, vol. 27, 1979, pp. 113–12
15. E. Petajan, "Automatic lipreading to enhance speech recognition," in IEEE Global Telecommunications Conference, (Atlanta, GA, USA), pp. 265–272, IEEE, 1984.
16. Sagheer, A., Tsuruta, N., Taniguchi, R. I. & Maeda, S. (2006). Appearance feature extractionversus image transform-based approach for visual speech recognition, International Journal of Computational Intelligence and Applications, Vol. 6, pp. 101–122.
17. Lucey, P., & Sridharan, S. (2008). A Visual Front-End for a Continuous Pose-Invariant Lipreading System, Proceedings of the 2nd International Conference on Signal Processing and Communication Systems, 15-17 December 2008, Australia, Queensland, Gold Coast.