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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 8, August 2013

Study of Audio-Visual Feature in Noisy Background

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: In recent development of technology the world is been surrounded by the latest technology which are based on man machine interaction. The key role is played by mobile phone and video conferencing in which quality of speech is the main criteria. Here in this paper we have reported the enhancement of Audio Speech recognition with the help of visual feature in noisy background. For extracting visual feature DCT along with the Wavelet is used and audio feature are extracted using MFCC feature. Further the results are reported at different signal to noise ratio (SNR) i.e. ranging between 20 dB to -5 dB using AWGN.

Keywords: DCT, MFCC, SNR, Speech

I. INTRODUCTION

Today technology demands for a powerful computing device which can easily utilized in a commercial way by the people which can lead to increase in the productivity as well as in recreation can only be realized with proper human–machine communication. These ideas produce an Automatic speech recognition system the most important step toward natural human–machine interaction. Motivation led by the researcher in the field of automatic speech recognition produce remarkable results, leading to many exciting expectations as well as new challenges. On the other hand speech is the primary, and the most convenient, means of communication between people. The significance role toward development of the technologies was become important during World War II, toward technological curiosity to build machines to mimic humans or the desire to automate work with machines. The advent of powerful computing devices further gives hope to this relentless pursuit, particularly in the past few years.

One of application which is primary used for the security purpose is the Speaker verification. Speaker verification (SV) is a technology which is used to authenticate persons from their voice samples. For this technique the commonly used feature for audio is Mel-frequency cepstral coefficients (MFCC). The main applications of the SV technology are in person authentication and in forensic science.

On the other hand with the recent growth in the wireless telecommunication, many of these applications are now accessed through mobile phones. In such scenario there are chances of large variation during the verification process due to handset devices and environmental conditions in addition to channel variability. So the attempt are been made by the researcher for minimizing the effect due to presence of background noise. Motivation is to design the robust system which can mimic the effect of noise. Here in this paper one of the techniques used to minimize such effect is to use the visual feature along with the audio feature to enhance the performance of the system.

II. HISTORICAL RESEARCH FOR HUMAN-MACHINE INTERACTION

The goal of speech recognition is for a machine to be able to "hear," understand," and "act upon" spoken information. During the initial stages attempts to design systems for automatic speech recognition were mostly guided by the theory of acoustic-phonetics, which clearly explain the elements of speech and how they are realized to form a spoken language.

The earliest speech recognition systems were first attempted in the early 1950s at Bell Laboratories, where researcher developed an isolated digit Recognition system for a single speaker [1].

More emphasis was given in the 1960s by the Japanese laboratories, for building a special hardware. Among them one was vowel recognizer of Suzuki and Nakata of the Radio Research Lab in Tokyo [2]. Major research was carried out in the Bell Laboratories' were approach was to take the concept of keyword spotting as a primitive form of speech understanding [3]. This application was used to detect prescribed words or phrases of particular significance, while neglecting those nonessential portions of the utterance. This is owing to the need to accommodate talkers who often

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 8, August 2013

prefer to speak natural sentences rather than rigid command words.

But later on after too much research in the field of automatic speech recognition it was reported that due to adverse environment condition the recognition rate is affect to large extent. Even it was reported in [4] that even there is clean background condition still the recognition rate is degraded. These two approaches had a profound influence of adding the visual feature along with the audio feature to improve overall human–machine speech communication technology in the past decades.

References [5] have reported the literature review of audio-visual speech recognition and also result are evaluated for the clean environment. Here the extension work of [5] is reported in which the performance of AVASR is evaluated for of the noisy audio signal at different SNR.

III. EXPERIMENTAL APPROACH

The figure 1.1 shows the basic procedure for Audio-Visual Speech Recognition. Has it been reported in [5] that original video signal is broken into two part i.e. i.e. audio and frame (image).

From fig 1.1 the two modalities are evaluated separately and then they are merged together to form as a single modality.

Figure 1.2 shows the procedure for extracting the audio feature. In Audio front end the pre-processing of the signal is

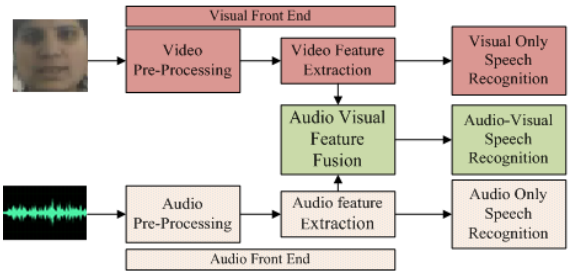


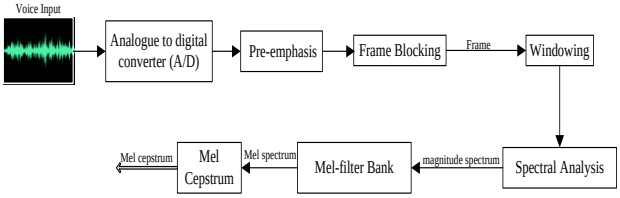
Figure 1.1 Basic procedures for AVASR

done to remove the redundancy if any present and thereafter the audio feature are extracted.

Audio feature are extracted using MFCC and then audio only recognition is performed, but here little modification is done at the audio level. In audio signal a noise is added and then the features are extracted so that the performance of noisy audio speech signal can be analysed.

In similar fashion the visual feature are evaluated with the help of video front end in which the frame are extracted and the pre-processing is done in the same way as done for audio.

Here one assumption is taken that the noise that is injected in the audio signal is not affecting our visual feature since the audio feature and visual feature are complementary to each other.



Finally the features evaluated are integrated together for checking the performance of the designed AVASR. 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 using.

IV. RESULT ANALYSIS

Basically our research work is divided into three sections: the design of the phoneme based recognizer followed by the design of the viseme only based recognizer and then the integration of audio and visual modes of speech to design the phoneme-viseme based recognizer. This section includes the detail of the work performed.

The Hindi database was used in which 5 speakers who have recorded 10 sentences each i.e. total 50 sentences. During the recording the background conditions of all the speakers are kept almost the same. Hindi language has been chosen as the basis of our work. This is because Hindi as a language is far more advanced than English. It is written as it is spoken and even has larger no. of phonemes as compared to English. Audio recordings were sampled at 48 kHz.

Reference [1] has evaluated the result for clean audio signal as shown in figure 1.3.

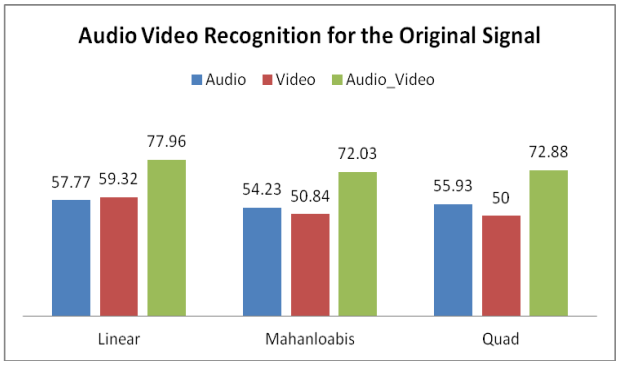


Figure 1.3 Shows the Audio-Video Recognition of original Signal.

On the other hand figure 1.4 shows the Audio-video recognition for 20 dB noisy signal.

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 8, August 2013

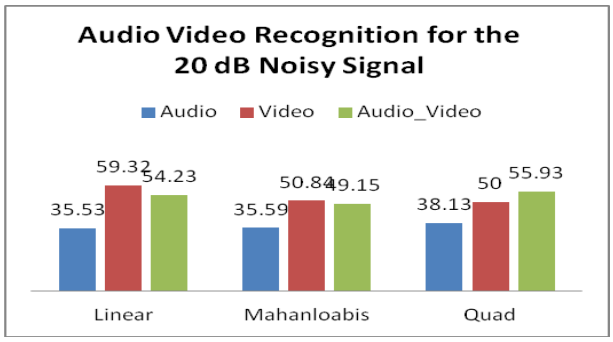


Figure 1.4 Shows the Audio-Video Recognition at 20 dB Noisy Signal.

Similarly figure 1.5 shows the Audio-video recognition for 10 dB noisy signal.

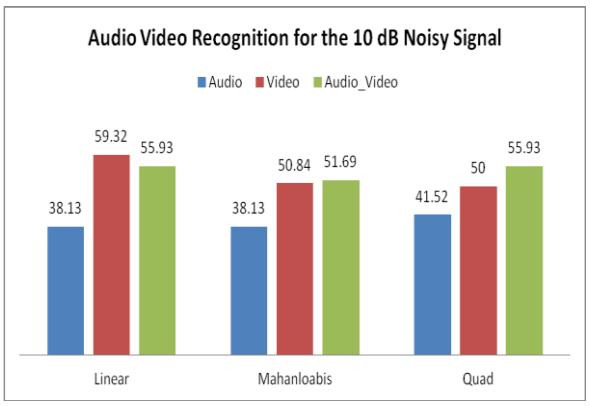


Figure 1.5 Shows the Audio-Video Recognition at 10 dB Noisy Signal.

Similarly figure 1.6 shows the Audio-video recognition for 0 dB noisy signal.

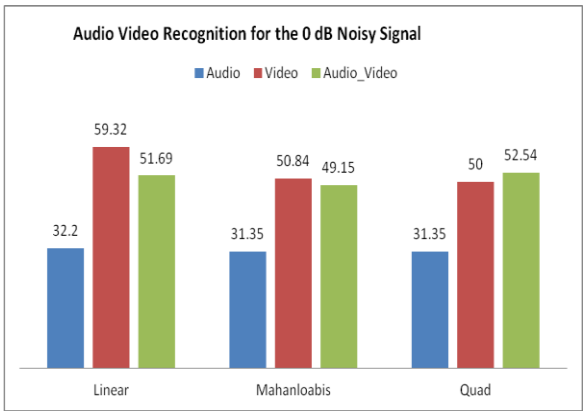


Figure 1.6 Shows the Audio-Video Recognition at 0 dB Noisy Signal.

Similarly figure 1.7 shows the Audio-video recognition for -5dB noisy signal.

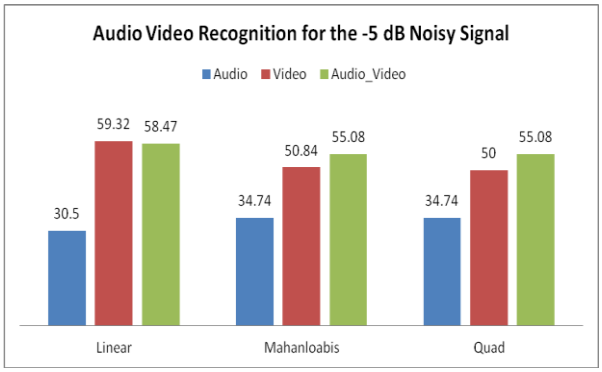


Figure 1.5 Shows the Audio-Video Recognition at -5 dB Noisy Signal.

V. CONCLUSION

As the result obtained for the percentage recognition of viseme for Audio only recognition (A),Video Only recognition (B), Audio Video using 49 feature i.e.(13 MFCC plus 36 DCT) feature(C), it can be easily concluded that the AVASR have a significant role over audio only recognition when the presence of noise degraded the audio quality.

It can be easily being shown in Figure 1.3 for a clean audio signal that 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.

Also in Figure 1.4 when 20 dB noises is injected in audio signal and the same procedure was carried out, recognition rate is increased by 18.7%, 13.56% and 17.80% for Linear, Mahalanobis and Quadratic classifier respectively for the optimum value.

In Figure 1.5 when 10dB noise is injected in audio signal and the same procedure was carried out, recognition rate is increased by 17.8%, 13.56% and 14.41% for Linear, Mahalanobis and Quadratic classifier respectively for the optimum value.

In Figure 1.6 when 0dB noise is injected in audio signal and the same procedure was carried out, recognition rate is increased by 19.49%, 17.88% and 21.19% for Linear, Mahalanobis and Quadratic classifier respectively for the optimum value.

In Figure 1.7 when -5dB noise is injected in audio signal and the same procedure was carried out, recognition rate is increased by 27.97%, 20.34% and 20.34% for Linear, Mahalanobis and Quadratic classifier respectively for the optimum value.

Finally it can be conclude from theory that when the noise degraded the quality of the audio signal then the presence of the visual information can be used to improve the quality of the overall system.

Hence the combinations of these two modality i.e. audio feature and visual feature have shown the significantly much

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 8, August 2013

improvement over single modality. Hence the purposes of combining these two modality have result in the design of the robustness system. Hence AVASR has given the robustness result for the system.

REFERENCES

[1] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," J. Acoust. Soc. Amer., vol. 24, no. 6, pp. 637–642, 1952.

[2] H. F. Olson and H. Belar, "Phonetic typewriter," J. Acoust. Soc. Amer., vol. 28, no. 6, pp. 1072–1081, 1956.

[3] J. G. Wilpon, L. R. Rabiner, C. H. Lee, and E. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden Markov models," IEEE Trans. Acoust., Speech, Signal Processing, vol. 38, pp. 1870–1878, Nov. 1990.

[4] Ashish Verma, Tanveer Faruquie, Chalapathy Neti, Sankar Basu. Late Integration in Audio-Visual Continuous Speech Recognition. IBM Solutions Research Center, New Delhi, India.

[5] Priyanka Sharma, Parul Dihulia, Vikas Gupta."Enhancement of ASR using viseme clue", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 1, April 2013.