Audio and Audiovisual Speech Recognition for Noisy Acoustic Conditions

Amaresh P. Kandagal and V. Udayashankara

Sri Siddhartha Institute of Technology,
Sri Siddhartha Academy of Higher Education,
Tumkur, India.
amareshpk@gmail.com

Sri Jayachamarajendra College of Engineering,
Mysore, India.
v_udayashankara@sjce.ac.in

Abstract

Speech Recognition (SR) is a primary feature for Human-Computer Interactions (HCI). ASR system is performing with high recognition rate only in acoustic controlled environments. It is very much significant and challenges to improve the performance of ASR systems in noisy conditions. The accuracy of recognition system can be improved with an additional source of complementary information along with acoustic features. In this paper, the performance of ASR system for various acoustic noises at different signal-to-noise ratio’s (SNR). By Otsu threshold algorithm lip geometric features are estimated. Acoustic features are estimated using Mel-frequency co-efficient method. Feature level fusion is carried out to develop Audio-visual automatic speech recognition (AV-ASR) system. The system performance is evaluated with various acoustic noises and different SNR’s. The recognition rate of 97.808% is noted for clean acoustic models those are tested with clean speech data. However, the same acoustic models that are examined with different environmental noisy speech test data with SNR 20dB and 5dB the average recognition rate is 91.45% and 62.902% respectively. The noisy audiovisual models are evaluated under various noisy acoustic conditions for SNR 20dB and 5dB the average recognition rate is recorded 96.746% and 89.012% respectively.

Key Words: ASR, AV-ASR, bimodal, HMM, Otsu, MFCC, lip reading.
1. Introduction

Automatic Speech Recognition (ASR) systems have undergone a large amount progression over the last few decades. Modern speech recognition solutions are able to produce reliable accuracy; however, in various real-world conditions, the encroachment of acoustic noise severely impact the performance of the system [1,2,3]. ASR system is a vital component for human-computer interfaces and intuitive to interact with smart devices such as smartphones [4,5]. We carry these devices in noisy acoustic environments such as while travelling use in vehicles, at offices, and many public places (airports, in railways, railway stations, and shopping malls etc). It is very much important that the performance of ASR solutions should be high at ambient noise levels. To overcome this, introduce another modality which will not affect to acoustic noise that is visual features (lip tracking) to complement the speech information [6, 7, 8, 9]. When both acoustic and visual information is available, it is obvious that humans tend to use both the modalities for HCI. The conception of humans applies bimodal perception is presented by the ‘McGurk effect’, or as ‘hearing lips and seeing voices’ [10]. The initial work on Automatic Lip Reading (ALR) appeared in 1984 when Petajan [11] investigated the use of lip features to enhance speech recognition by creating a bimodal (audio-visual) speech recognition system [12, 13, 14].

This audio-visual perception yields a rich combination of information and is utilized in the development of speech recognition system. From speaker’s mouth region visual features extracted i.e silent speech features have more potential information for boosting the accuracy of a recognition system for noisy environmental conditions [15]. However, combining audio and visual streams is essential for designing a robust audiovisual recognizer [16]. Comparative analysis is done using and without using visual features to recognize the speech under various environments.

Figure 1 shows the general architecture of the feature fusion based audiovisual speech recognition system. As the described earlier AVASR system uses acoustic information in addition to the visual speech information. Acoustic and visual features can be concatenated at feature level (feature fusion is also known as early integration) or decision level (decision fusion is also known as late integration) [17].

The visual features merged at the feature level, and this collective audiovisual observation sequence is passed to recognition unit (i.e decision classifier). The straightforward and simple approach of feature fusion is a concatenation of audio and visual features [18]. The hierarchical linear discriminant feature extraction approach feature weighting and audio feature enhancement methods are depicted in [13,19,20,21]. Feature fusion based methods are depicted in capability to explicitly model the relative reliability of either feature channel. It is significant as the reliability of any stream may alter comprehensively even
within the period of an utterance due to instantaneous or constant channel degradations or background noise [22].

The decision fusion method presumes independence between the audio and video observation sequences and integrates the results of respective stream classifier. In contrast to early integration approach, late integration offers a technique for modeling the reliabilities of respective feature stream by using individual classifier for video and audio channels. The decision fusion methods combine parallel classifier architecture with multi stream HMM (MSHMM) with fixed or adaptive combination weights [23, 24, 25].

![Figure 1: Architecture of the Feature Fusion Based Audiovisual Speech Recognition System](image)

The rest of the paper is organized into different sections. Section 2 presents literature survey on AVSR system, section 3 describes database creation, section 4 explains the methodology of the proposed system, section 5 describes feature integration, section 6 presents experimental results and Section 7 discusses on conclusions and future enhancements of the proposed work.

2. Related Work

Subramanya Amarnag et.al [8] proposes, Coupled Hidden Markov Model (CHMM), audio SNR and noise type for the purposes of audio-visual integration. The experimental results indicate that the CHMM system trained using the knowledge of the SNR and the noise type outperforms the
conventional multi-stream Hidden Markov Models (MSHMM) by as much as 8% for an audio SNR of 6db. Knowledge of the noise type improves the recognition scores; it would be useful in applications, where prior knowledge of the application of the recognizer is available.

Kwanchiva Thangthai et al. [26] investigated the development of AVASR system for noisy environments by using deep neural networks (DNNs). The preliminary experiments are conducted by using the Kaldi toolkit [37]. The performance of the system is compared against conventional hidden Markov models. The accuracy of the system is measured with different signal to noise ratio levels of babble noise ranging from 20dB to -20dB. The authors claim that combining the visual features by early integration is efficient for the signal to noise ratio at 5dB and above.

M. Z. Ibrahim et al. [27] uses feature fusion based AVSR system to extract geometry lip from the region of mouth by making use of convex hull, border following and the filter of skin color and Hidden Markov Model for classification. This approach is compared with the conventional system of audio while under the operation of simulating conditions of ambient noise which affects the phrases spoken. The approach improves the accuracy of speech recognition when compared with an approach that uses only audio.

Anusha Kistapuram et al. [28] has adopted grouping of feature fusion methods and decision fusion methods. The first method is dependent on the single classifier of combined audio and visual feature vector. Concerning the algorithm of decision fusion, it uses outputs of two single modalities which are individual audio and visual classifier only for recognition of audiovisual speech. This can be obtained by linear combination of classes of log likelihood conditional observations of two classifiers to joint of audio classification score and visual classification score by the use of appropriate weights which captures single modality classifier reliability.

Tian Gan et al. [29] present an approach to articulatory features based audiovisual speech recognition in which video frames are tried to combine. Among independent classifiers, the best N outputs are combined to determine the advantages and are compared with phone-based recognizer to show better results. Abstract articulatory classes are modelled by using HMMs that are extracted parallel from particular feature tuples and the speech signal. These tuples are mapped to phones and stream of the phone is generated. Finally, this stream is mapped to the meaningful words from lexical search.

3. Database and Experimental Setup

In this work, in-house database created. Digital camera with the 5MP camera is considered for recording videos of resolution 320x240 with 30 frames per second and sampling frequency of 44100Hz for audio. Recordings were captured from 63 subjects: 36 female, and 27 male. The database is created for
15 Kannada words and 10 sentences (combination of 4 to 9 words). Corpus collection conducted in two different sessions. On every session, each word is captured for 2 occurrences from respective speakers.

4. Methodology

Visual Feature Extraction

Lip segmentation methods are broadly categorized into color-based techniques and model-based techniques. Figure 2 illustrates color based lip feature extraction modules.

Choosing appropriate color stream plays a significant role, in the course of prolific feature extraction. Colored images (composed of RGB) are computationally expensive compared to gray scaled image frames. Therefore these key frame images are converted to a grayscale format to reduce the computation complexity. Grey scale transformation for lip map generation \(Im\) is computed as

\[
Im = a \cdot R - b \cdot G + c \cdot B
\]  

Where \(a=0.2, b=0.6\) and \(c=0.3\)

\[
Im_{\text{norm}} = \frac{(Im - Im_{\text{min}})}{Im_{\text{max}} - Im_{\text{min}}}
\]

The Otsu’s algorithm implementation is a gray threshold function in MATLAB® is given by [10].

\[
\sigma^2_{\text{intra}} = x_1 \sigma^2_1 + x_2 \sigma^2_2
\]
Where $\sigma_i^2$ and $\sigma_T^2$ are variance of class $i$ and variance of the total image respectively. $x_i$ is the probability of a pixel and $m_i$ mean of class $i$.

The ratio of lip height by width estimates lip motion angle $\theta$. As lip shape is symmetrical in nature by measuring the midpoint of width; we estimate the upper and lower lip height.

$$\theta = \frac{(H1 + H2)}{W}$$

Where $W$ width of the lip, $H1$ and $H2$ are upper and lower lip height respectively [30].

**Acoustic Feature Extraction**

The acoustic signal is removed from the unwanted background noise and is enhanced by spectral subtraction algorithm. The boundary detection method is carried out by zero crossing rates.

The conventional MFCC algorithm is applied to extract the prominent features. The proposed system is developed by making a variation in framing level and overlapping. Parameters are estimated with the 35% of overlapping of frames for every 25 milliseconds of a signal. Feature vector was of thirty-nine, with twelve Mel cepstrum plus log energy and first and second order derivative coefficients i.e delta and acceleration coefficients [31, 32].

## 5. Feature Integration

The computed (audio and visual) features from different modalities integrated intelligently to make use of them effectively in the classifier module. As discussed earlier feature fusion and decision fusion are two approaches to work with different kind of features. In this presented paper, feature fusion approach is applied to develop the AVASR system. The acoustic and visual are in different frame rates, 44100Hz and 30Hz for audio and video respectively. Therefore linear interpolation is applied to up sample the video features rate to accomplish feature integration. To realize feature fusion, we basically concatenate the acoustic and visual features and subsequently by applying a classification algorithm to recognize the observation sequence. The thirty-nine acoustic features and fifteen visual features are integrated to produce a fifty-four feature audio-visual single vector.

The early integration model is based on a traditional HMM classifier on the concatenated vector of the audio and visual features $O_t = [O^A_t, O^V_t]$ where $O^A_t$ and $O^V_t$ denote the audio- and visual-only feature vectors at time instant $t$. 

\[
\begin{align*}
\sigma_{\text{Inter}}^2 &= x_1x_2(m_2 - m_1)^2 \quad \text{(4)} \\
\sigma_T^2 &= \sigma_{\text{Inter}}^2 + \sigma_{\text{Intra}}^2 \quad \text{(5)} \\
\eta &= \frac{\sigma_{\text{Inter}}^2}{\sigma_T^2} \quad 0 \leq \eta \leq 1 \quad \text{(6)}
\end{align*}
\]
6. Classification

Hidden Markov Models [33, 34] are the robust and most popular classifier algorithms for speech recognition system and these are a well-liked and proven to be a highly reliable classifier for the statistical modeling of audible speech [35].

Seven-state left-to-right audio-visual word models are created for every Markov state being modeled with diagonal covariance of 5-Gaussian mixers exclusively. Baum-Welch algorithm is applied to create acoustic models and audiovisual models and the Viterbi algorithm is applied to recognize unknown speech or audiovisual observation sequence [33, 36].

7. Experimental Results and Discussion

We have conducted the experiments to evaluate the proposed system in various noisy acoustic conditions. Noise signals are chosen to signify the most likely application circumstances for telecommunication terminals.

Acoustic noises are recorded from various places as Crowd of people (Babble), inside and outside the vehicle (Car), Street, Factory, Subway, Bus station and railway stations.

Noise signals are moderately stationary such as e.g. the car noise. Others comprise with non-stationary segments such as the recordings on streets, airport, railway stations etc.

Observations from the long-term spectra the majority part of the energy of the signal concentrates in the low-frequency region. With reference to the spectral point of view, most of the acoustic noises appear to be pretty alike even though they are recorded in different environmental acoustic conditions. The acoustic noise is combined to the acoustic corpus at the signal to noise ratios of 20, 15, 10, 5, 0 and –5dB.

Experiments are conducted with and without using visual features for clean and noisy speech data.

The advantage of building acoustic models with clean speech dataset itself is modeling of speech not containing any kind of acoustic noise distortions and these models are suitable to represent all available acoustic information. Reasonably this kind of ASR systems performs highest recognition rate on testing with clean data itself.

However, these models do not hold any kind of information about possible acoustic distortions. Hence multi-condition models are built by using distorted speech signals as training data.
The system performance results are tabulated for speaker independent ASR, VSR and AVSR system. Out of 36 female and 27 male speakers, 26 female and 20 male speaker corpuses is applied for training purpose and the leftover data is used for validation. The results are provided for clean and noisy acoustic models. Table 1 represents the results for audio-only speech recognition for clean acoustic data models. The recognition rate of 97.808% is noted, when clean acoustic models are tested with clean speech data. However, for the same acoustic models that are examined with different environmental noisy speech test data with signal-to-noise ratio (SNR) 20dB and 5dB the average recognition rate is 91.45% and 62.902% respectively.

In addition, system performance is evaluated for noisy acoustic models with clean speech testing data the average recognition rate is 97.62% and furthermore with noisy speech testing data for 20dB and 5dB SNR the average recognition rate is noted 95.98% and 85.11% respectively. These results are tabulated in Table 2.

Two sets of audiovisual models are trained i.e by the fusion of visual information along with clean and noisy speech data the corresponding clean and noisy audiovisual models are generated. The experimental results for clean audiovisual models are tabulated in Table-3 and the noisy audiovisual models are evaluated under various noisy acoustic conditions for SNR 20dB and 5dB the average recognition rate is 96.746% and 89.012% respectively. These results are tabulated in Table 4.

As per the observation from the experimental results, the performance of ASR system degrades with noisy acoustic environmental conditions. However, AV-ASR system performance is significantly improved and acceptable for noisy acoustic conditions. Figure 3 represents the ASR, VSR and AV-ASR performance with reference to average overall noises. The VSR system performance is not affected the noisy acoustic conditions.

8. Conclusions

The results show the visual features contains information regarding that boosts the performance of the system under various noisy acoustic conditions. The performances of the systems of conventional audio-only and audiovisual results were compared under various noises and a range of different SNRs. The recognition rate of 97.808% is recorded for clean acoustic models are tested with clean speech data. However, the same acoustic models that are examined with different environmental noisy speech test data with SNR 20dB and 5dB the average recognition rate is 91.45% and 62.902% respectively. The noisy audiovisual models are evaluated under various noisy acoustic conditions for SNR 20dB and 5dB the average recognition rate is recorded 96.746% and 89.012% respectively. It is observed that multi-condition models improve recognition rate. The performances of the system depreciate by diminishing signal-to-noise ratio.
## Table 1: Speech Recognition with Clean Acoustic Models

<table>
<thead>
<tr>
<th>SNR (in decibels)</th>
<th>Babble</th>
<th>Car</th>
<th>Horn</th>
<th>Factory</th>
<th>Subway</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>98.05</td>
<td>97.28</td>
<td>98.35</td>
<td>97.54</td>
<td>97.82</td>
<td>97.808</td>
</tr>
<tr>
<td>20</td>
<td>92.46</td>
<td>91.55</td>
<td>91.8</td>
<td>90.42</td>
<td>91.05</td>
<td>91.456</td>
</tr>
<tr>
<td>15</td>
<td>89.5</td>
<td>86.45</td>
<td>87.02</td>
<td>83.5</td>
<td>84.7</td>
<td>86.234</td>
</tr>
<tr>
<td>10</td>
<td>78.38</td>
<td>75.08</td>
<td>73.15</td>
<td>72.35</td>
<td>74.05</td>
<td>74.602</td>
</tr>
<tr>
<td>5</td>
<td>60.65</td>
<td>63.45</td>
<td>64.21</td>
<td>63.05</td>
<td>63.15</td>
<td>62.902</td>
</tr>
<tr>
<td>0</td>
<td>51.33</td>
<td>48.05</td>
<td>50.43</td>
<td>51.08</td>
<td>52.45</td>
<td>50.668</td>
</tr>
<tr>
<td>-5</td>
<td>20.63</td>
<td>23.75</td>
<td>22.38</td>
<td>26.9</td>
<td>28.9</td>
<td>24.512</td>
</tr>
</tbody>
</table>

## Table 2: Speech Recognition with Noisy Acoustic Models

<table>
<thead>
<tr>
<th>SNR (in decibels)</th>
<th>Babble</th>
<th>Car</th>
<th>Horn</th>
<th>Factory</th>
<th>Subway</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>97.7</td>
<td>97.2</td>
<td>98.09</td>
<td>97.45</td>
<td>97.68</td>
<td>97.624</td>
</tr>
<tr>
<td>20</td>
<td>95.54</td>
<td>96.43</td>
<td>94.73</td>
<td>96.14</td>
<td>97.1</td>
<td>95.988</td>
</tr>
<tr>
<td>15</td>
<td>93.75</td>
<td>94.64</td>
<td>92.15</td>
<td>94.01</td>
<td>95.35</td>
<td>93.98</td>
</tr>
<tr>
<td>10</td>
<td>88.44</td>
<td>90.82</td>
<td>87.45</td>
<td>90.31</td>
<td>93.6</td>
<td>90.124</td>
</tr>
<tr>
<td>5</td>
<td>83.36</td>
<td>86.35</td>
<td>82.18</td>
<td>83.9</td>
<td>89.76</td>
<td>85.11</td>
</tr>
<tr>
<td>0</td>
<td>66.3</td>
<td>68.65</td>
<td>67.35</td>
<td>63.68</td>
<td>67.9</td>
<td>66.776</td>
</tr>
<tr>
<td>-5</td>
<td>30.63</td>
<td>37.67</td>
<td>30.85</td>
<td>40.34</td>
<td>45.3</td>
<td>36.958</td>
</tr>
</tbody>
</table>

## Table 3: Audiovisual Speech Recognition

(Audiovisual Models Trained with Clean Speech Data)

<table>
<thead>
<tr>
<th>SNR (in decibels)</th>
<th>Babble</th>
<th>Car</th>
<th>Horn</th>
<th>Factory</th>
<th>Subway</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>98.54</td>
<td>98.02</td>
<td>98.16</td>
<td>98.14</td>
<td>98.06</td>
<td>98.184</td>
</tr>
<tr>
<td>20</td>
<td>94.16</td>
<td>95.15</td>
<td>93.25</td>
<td>91.24</td>
<td>92.05</td>
<td>93.17</td>
</tr>
<tr>
<td>15</td>
<td>90.5</td>
<td>89.85</td>
<td>88.22</td>
<td>84.03</td>
<td>85.27</td>
<td>87.574</td>
</tr>
<tr>
<td>10</td>
<td>83.18</td>
<td>79.77</td>
<td>76.9</td>
<td>78.55</td>
<td>77.94</td>
<td>79.268</td>
</tr>
<tr>
<td>5</td>
<td>65.15</td>
<td>67.15</td>
<td>69.09</td>
<td>68.88</td>
<td>67.25</td>
<td>67.504</td>
</tr>
<tr>
<td>0</td>
<td>54.28</td>
<td>52.96</td>
<td>51.09</td>
<td>53.78</td>
<td>53.12</td>
<td>53.046</td>
</tr>
<tr>
<td>-5</td>
<td>26.33</td>
<td>29.15</td>
<td>30.13</td>
<td>29.77</td>
<td>29.29</td>
<td>28.934</td>
</tr>
</tbody>
</table>

## Table 4: Audiovisual Speech Recognition

(Audiovisual Models Trained with Noisy Speech Data)

<table>
<thead>
<tr>
<th>SNR (in decibels)</th>
<th>Babble</th>
<th>Car</th>
<th>Horn</th>
<th>Factory</th>
<th>Subway</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>97.85</td>
<td>97.82</td>
<td>97.65</td>
<td>97.55</td>
<td>97.8</td>
<td>97.734</td>
</tr>
<tr>
<td>20</td>
<td>96.15</td>
<td>96.5</td>
<td>96.35</td>
<td>97.05</td>
<td>97.68</td>
<td>96.746</td>
</tr>
<tr>
<td>15</td>
<td>95.05</td>
<td>95.05</td>
<td>94.5</td>
<td>95.8</td>
<td>95.88</td>
<td>95.256</td>
</tr>
<tr>
<td>10</td>
<td>93.65</td>
<td>93.25</td>
<td>91.26</td>
<td>92.15</td>
<td>94.9</td>
<td>93.042</td>
</tr>
<tr>
<td>5</td>
<td>91</td>
<td>89.2</td>
<td>86.15</td>
<td>88.64</td>
<td>90.07</td>
<td>89.012</td>
</tr>
<tr>
<td>0</td>
<td>72.65</td>
<td>71.5</td>
<td>69.95</td>
<td>68.75</td>
<td>72.45</td>
<td>71.06</td>
</tr>
<tr>
<td>-5</td>
<td>54.25</td>
<td>52.45</td>
<td>48.85</td>
<td>51.56</td>
<td>54.05</td>
<td>52.232</td>
</tr>
</tbody>
</table>
The degradation does not considerably diverge for the dissimilar acoustic noises.

As a future enhancement, the proposed system can be extended to improve the visual feature selection and performance evaluation with decision feature fusion and conducting experiments with different classifiers such as neural networks.

Acknowledgements

We thank Dr. Anusuya M A, Professor, Department of Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, JSS Science and Technology University, Mysuru, for her significant support in understanding and implementing this presented paper.

References


