1. Introduction

The human capacity to perceive speech through visual cues, the ability to lipread is well known. In an historical review given by French-St George [01] and a 17th century example describing speechreading is Bulwer [02]. Today it is clear that it is not only hearing-impaired people that benefit from, and use, the visual aspect of speech. Even those with normal hearing are better able to understand speech if they can see the talker, Reisberg [03]. Visual enhancement is even more pronounced when the auditory signal is degraded and noisy. Neely [04] found an increase in intelligibility of 20% with the addition of vision and others have reported similar findings [05, 06, 07].

A powerful demonstration of the auditory-visual nature of speech comes from the McGurk effect [08, 09, 10]. When presented with conflicting auditory and visual symbols, under certain conditions, the perceiver ‘hear’ something other than spoken word due to conflicting visual stimulus of spoken sound. An example fusion illusion occurs when an auditory /ba/ is synchronized with a visual /ga/. Thus is it usually perceived as /da/, different from either the visual or auditory stimuli. Therefore it occurs even when the perceiver is aware of the illusion, and has been demonstrated in four month old infants [11] as well as adults.

Another demonstration comes from experiments with auditory-visual asynchrony. McGrath [12] found an audio lead of less than 80ms or lag of less than 140 ms could not be detected during speech. However, if the audio was delayed more than 160ms it is no longer contributed useful information and the subject reported identification accuracy down to the visual-only level. In practice, the delays up-to 40ms are
acceptable. It is worth noting that in television broadcasting terms this represents a single PAL (Phase alternate Line) frame of asynchrony. Evidently any audio-visual speech recording must take care to maintain audio-visual synchronization. A study by Kuhl [13] found that even five month old infants prefer the face of talker with synchronized audio to one without.

2. Literature Survey

The work on automatic speechreading was introduced by Eric Petjan [14] at AT&T Bell labs, where the work was concentrated on improvement on automatic lipreading system which incorporates Dynamic Time Warping (DTW), and Vector Quantization (VQ) method applied on alphabets and digits using four speakers. The recognition was restricted to isolated utterances and was speaker dependent. The speaker faced solid state camera which sent digitized video data to microcomputer system with special purpose video processing hardware to perform windowing, grayscale thresholding and contour coding at video rate. The video data was sampled during utterances and then reduced to a template consisting binary image sequence of mouth. Registration (Realignment) was done by considering nostrils from frame to frame assuming very small variations in distance between the nostril and mouth. While recognition each sample utterances was matched against all other sample utterances using dynamic programming to optimize the match. The average distance between given sample and the sample of the given letter were computed for each letter. The letter was then ranked as to distance to the given sample and the error term was recorded if first ranked letter was not the same as the sample letter.

Christoph Bregler [15, 16] had worked on how recognition performance in automated speech perception can be significantly improved by additional lipreading also called as speechreading where they introduced extension to existing Multi-State Time Delayed Neural Network (MSTDNN) architecture for handling both the modalities that is acoustics and visual sensor input. The input video images are grabbed in real-time (30 full frames per second) with 256x256 pixel image with 8 bit grey level information per pixel. This squared region will cover full face of the speaker similarly acoustic data is sampled at 16 KHz rate and 12 bit resolution, in continuation timestamps were saved to obtained correct synchronization between audio and video signals. The database formed using this process was consists of 114 and 350 letter sequences spelled by two male speakers. They consist of names and random sequences. The first data set was split into 75 training and 39 test sequences.
Second data was split into 200 training and 150 test sequences. The back propagation was applied with learning rate 0.05 and momentum of 0.5 also different error functions to compute the error derivatives. They have extended this work by introducing more general task of learning constraint surfaces. They described simple but powerful architecture for learning and manipulating non linear surface from data. They have demonstrated the technique on low dimensional synthetic surfaces and compared it with the nearest neighbor approach. They have shown its utility in learning the space of lip images in a system for improving speech recognition by lipreading. This learned surface is used to improve the visual tracking performance during the recognition. Similar works have been done by Yuhas et.al in 1993, focusing on neural network for vowel recognition and worked on static images [17].

Michel Brooke [18] has proposed that visible facial gestures could be exploited to enhance speech intelligibility in automatic system and handling this voluminous data that is images of talkers face implies some form of data compression. Therefore rather than using conventional feature extraction approach, image coding and compression can be achieved using data driven, statistically-oriented techniques such as Artificial Neural Network (ANNs) or Principal Component Analysis (PCA) can be used. The major issue addressed by him was the combination of audio and visual data so that the best use can be made of the two modalities together. The perceptual experiments may offer guidance on suitable machine architecture, many of which currently using Hidden Markov Model (HMM).

Alen Goldchen and Eric Petjen [19] extended their earlier work [14] by introducing a continuous optical automatic speech recognizer that uses optical information from the oral-cavity shadow of a speaker. The system achieves 25.3% recognition on sentences having a perplexity of 150 without using any syntactic, semantic, acoustic or contextual guides. They explored 13, mostly dynamic, oral cavity features used for optical recognition, present phones that appear optically similar (Visemes) for the speakers and present recognition results for Hidden Markov Model (HMM) using Visemes, trisemes and generalized trisemes have been reported.

Paul Duchnowski [20] in their work on Towards movement invariant automatic lipreading and speech recognition, they focuses on development of speech recognition system incorporating automatic lipreading while allowing user freedom of movement within a room, whereas the acquisition process was continuous, automatic and without special markers but required the speaker to position himself in such as way
that his lips has to be appeared with in the defined window. The system was operating under two modes that is locating an arbitrary face and tracking the located face. The data was collected with image of 256x256 pixels at 30 frames per second and at 8 bit gray scale resolution as raw input. In order to reduce the amount of computations the hidden units have restricted as “receptive fields”. For calculating word accuracy they have trained recognizer on visual/ acoustic data from 200/1500 letter sequences from single speaker and tested on 30 sequences.

Juergen Luettin [21] worked on Visual speech recognition using active shape model and Hidden Markov Model (HMM). The shape of the mouth is modeled by an Active Shape Model (ASM) which is derived from the statistics of a training set and used to locate, track and parameterize the speaker’s lip movements. The extracted parameters representing the lip shape are modeled as continuous probability distributions and their temporal dependencies are modeled by Hidden Markov Models. The parameters describing the shape of the lips are extracted at each time frame and used as visual speech feature vectors. The parameters are invariant to scale, rotation, translation and illumination and can directly be used by the recognition network. The translation and rotation parameters are not used for recognition because they are unlikely to provide speech information. Much speech information is contained in the dynamics of the lip movements rather than the actual shape. Furthermore dynamics of lip movements might be less sensitive to linguistic variability. Therefore they performed some recognition tests by including temporal differences of each feature.

Rainer Stiefelhagen [22] has presented approach to lip tracking for lipreading. They proposed that instead of only tracking features on lips, we must track lips along with other facial features such as pupils and nostril. In this approach the face is first located in the image using a stochastic skin color model, the eyes, lip-corners and nostrils are then located and tracked inside the facial region. This approach can effectively improve the robustness of lip-tracking and simplify automatic detection and recovery of tracking failures. The feasibility of this approach was demonstrated by implementing a lip tracking system. The system has been tested by database that contains 900 image sequences of different speakers spelling words. The system has successfully extracted lip region from image sequences to obtain training data for the audio-visual speech recognition system. The system was also applied for extracting lip region from real-time video images to obtain visual input for an audio-visual
speech recognition system. On the test sequence they have achieved a reduction of number of frames with tracking failures by a factor of two using detection and prediction of outliers in the set of found features.

Roland Gocke [23, 24] has presented algorithm for robust and reliable automatic extraction of lip feature points for speechreading. The algorithm uses combination of color information in the image data and knowledge about the structure of mouth area to find certain features points on the inner lip contour. A new confidence measure qualifying how well the feature extraction process was introduced. A parameter set have been described by them for the shape of mouth on the basis of the position of feature points. While building lip features and parameter set for every speaker they have considered that every speaker has different lips, therefore they are interested in inner lip contour so that the personal characteristics, shape of the lips has minimal effect on the measurements. Furthermore, facial hair affects the visibility of outer lip contour.

Uwe Meier, Rainer Stiefelhagen [25] has extended their earlier work [22], by studying unrestricted lip rearing. They have introduced a top-down approach to automatically tract and extract lip regions. This technique makes it possible to acquire visual information in real-time without limiting user’s freedom of movements. They discussed normalization of algorithms to preprocess images for different lighting conditions. They have compared different visual preprocessing methods such as raw image, Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA). They demonstrated feasibility of their method by development of modular system for flexible Human Computer Interaction (HCI) via both visual and acoustic speech. This system has been evaluated under different noisy conditions such as white noise, music and mechanical noise. The experimental results indicate that the system can achieve upto 55% error reduction using addition visual information.

Anil K Jain [26] in the article on Statistical Pattern recognition: A Review, has addressed to different problem domains, applications, type of pattern allowed in problem domain with their pattern classes, article highlights on pattern recognition models in accordance with the Approach, Representation, Recognition function and Typical Criterion for building pattern recognition model. Feature selection method, Classification Method, Classifiers and their combination scheme, Error rates and error estimation methods, Clustering algorithms has been discussed.
Stephane Dupont and Juergen Luettin [27] has extended their earlier work [21] by building a system consist of three component a) visual module, b) acoustic module, and c) a sensor fusion module. The visual modules, locates and track the lip movements of a given speaker and extract relevant speech features. This task was performed with an appearance based lip model that is learned from example image. Visual speech features are represented by contour information of the lips and grey-level information of the mouth area. The acoustic module extracts noise-robust features from the audio signal and sensor fusion module is responsible for joint temporal modeling of the acoustic and visual feature system and is realized using multi stream Hidden Markov Model (HMM).

K.L. Sum [28] in their work on a new optimization procedure for extracting the point-based lip contour using active shape model. A 14-point ASM lip model is used to describe the lip contour. With the aid of fuzzy clustering analysis, a probability map of the color lip image is obtained and a region-based cost function is established. The new optimization procedure operates on the spatial domain (actual contour points) and all the points are pulled towards their desirable locations in each iteration. Hence, the lip contour evolution becomes better controlled and consequently fast convergence was achieved. In order to perform lip contour extraction automatically they have defined lip contour model using lip contour parameters. A 14-point ASM lip model was built using 200 lip images of 2 males and 2 females. The image size is 108×81 and in 24-bit true color (RGB). The ASM lip model has 26 bases, but only the first 6 significant bases are used in model.

A Capiler [29] in his work on lip detection and tracking has proposed algorithm based on Active shape model and Kalman Filtering in spatiotemporal framework where the mouth are deformed under constraints. He reported some limitations that lip motion can be very fast so that an acquisition rate of 25 frames per second is not enough and this can be resolved if camera with higher frame rate will be used. The proposed algorithm is made of three steps as

1. The preprocessing step which determines the correct model (OM or CM) and the initial shape (average shape or Kalman prediction) and automatically places it in the frame to be processed.

2. The processing step where the initial shape is deformed under constraints according to spatiotemporal energies and where the Kalman prediction is corrected according to the result of lip detection.
3. A post-processing step where the position of the mouth in the next frame is predicted by Kalman filtering. As lip detection is an iterative algorithm which requires a good initial estimates in order to converge quickly to obtain good result.

Joohum Lee and Jin Young Kim [30] have proposed an efficient method to reduce the amount of feature data for real-time automatic image transform based lipreading. Image transformed based approach obtaining a compressed representation of image pixel values of speaker’s mouth is reported to show superior lipreading performance. This approach produces many feature vector relevant to lip information, it requires much computation time for lipreading even when Principal Component Analysis (PCA) is applied. Therefore to reduce computational load efficiently they proposed an algorithm that utilizes symmetry of the lip. This method reduces the amount of required feature vector up to 51% as compared to original one. It also improves the recognition rates by compensating the variation of illumination.

Gerasimos Potamianos and Chalapathy Neti [31] have studied three aspects of designing appearance based visual features for automatic lipreading: a) the choice of the video region of interest on which image transform features are obtained, b) the extraction of speech discriminant features at each frame and c) use of temporal information to improve visual speech modeling. They proposed ROI to be formed in such a way that it includes speaker jaw’s and cheeks in addition to traditionally used mouth or lip region with respect to (b) and (c), they proposed, the use of two stage linear discriminant analysis, both within the frame, as well as across the large number of frames.

Sabri Gurbuz [32] has described the incorporation of visual lip tracking and lipreading algorithm that utilizes the affine invariant Fourier descriptors from parametric lip contours to improve the audio-visual speech recognition system. The audio-visual speech recognition system presented by them was based on parallel Hidden Markov Model (HMM) where joint decision, using an optimal decision rule is made after processing. This work described the extraction of affine invariant Fourier descriptors from parametric lip contour data. Finally they validated the use of optimal weight selection which is based on the noise type and signal-to-noise ratio (SNR) for joint audio-visual automatic speech recognition.

Ian Matthews [33] in their work on Extraction of Visual features of Lipreading, the major problem in this system is to generate visual features which are in enormous
quantity of data in video sequences and this is a problem that is common to all computer vision systems. Each video frame contains thousands of pixels from which a feature vector of about 10 and 100 elements must be extracted, ideally it was expected to select such features those are robust to different talkers, head poses, lighting conditions. The data of 10 talkers (5 Male, 5 Female) of the isolated words, with total 780 utterances. Each video utterance was digitized at quarter frame of 625 line frame video (376x288 @ 25 frames per second) in 8 bit gray scale recording. The feature extraction was done using low level statistical model to extract 60 features from 80x60 mouth images. The PCA have been used to identify orthogonal directions by their relative variance contribution

Kate Saenko [34], has discussed the approach to visual speech modeling based on articulatory features, which has potential benefits under visually challenging conditions. This idea was use to set for parallel Support Vector Machine (SVM) classifier to extract different articulatory attributes from the input images, and then combine their decisions to obtain higher-level units, such as Visemes or words. They evaluated their approach in preliminary experiments on a small audio-visual database, using several image noise conditions, and compare it to the standard Visemes based modeling approach.

Satoshi Tamura [35], have proposed a multi-modal speech recognition method using optical flow analysis for lip images. Optical flow is defined as the distribution of apparent velocities in the movement of brightness patterns in an image. Since the optical flow is computed without extracting lip contours and location, robust visual features can be obtained for lip movements. This method calculates two kinds of visual features sets in each frame. The first feature set consists of variances of vertical and horizontal components of optical flow vectors. These are useful for estimating silence/pause periods in noisy conditions since they represent movements of the speaker’s mouth. The second feature set consists of maximum and minimum values of integral of the optical flow. These are expected to be more effective than the first set since this feature has not only silence / pause information but also open / close status of the speaker’s mouth. Each of the features set is combined with an acoustic feature set in the framework of HMM based recognition.

M E Sargin [36], has used lip feature that has the highest correlation with audio features that has investigated. Audio features are selected as Mel Frequency Cepstral Coefficients (MFCC) of the audio signal. Three different lip features are considered
for the visual lip information, where these features are 2D Discrete Cosine Transform (DCT) coefficient of the intensity based image and optical flow vectors within the lip region, and the distances between pre-defined points on the lip contours which carries the lip shape information. Their study has demonstrated two techniques based on class conditional probability analysis and canonical correlation analysis to estimate and compare the correlations between audio feature and each lip feature. The lip feature, which has the highest correlation to audio features, is identified among the above lip features. Isolation of lip features, which are highly correlated with audio signal, can be used for audio-visual speech recognition, audio-visual lip synchronization and estimation of lip shapes using audio signal for visual synthesis.

Sadaoki Furui [37] has given overviews of robust architecture and modeling for spontaneous speech recognition and understanding. The topic includes acoustic and language modeling for spontaneous speech recognition, unsupervised adaptation of acoustic and language models, robust architecture for spoken dialogue systems, multi-modal speech recognition, and speech summarization.

Alaa Sagheer [38] has presented a visual speech feature representation approach that combines Hypercolumn Model (HCM) with Hidden Markov Model (HMM) to perform a complete lipreading system. They used HCM to extract visual speech features from input image. The extracted features are modeled by Gaussian distributions through using HMM. The proposed lip-reading system can work under varying lip position and sizes. All images were captured in a natural environment without using special lighting or lip markers. Experimental results were compared with SOM and DCT based system and HCM provides better performance than both the systems.

Xiaopeng Hong [39], in the work on A PCA based DCT features Extraction method for lipreading has presented a PCA based method to reduce the dimensionality of DCT coefficients for visual only lip-reading systems. A three-stage pixel based visual front end was adopted. First, DCT or block-based DCT features are extracted. Second, Principal Component Analysis (PCA) is applied for dimension reduction. Finally, all the feature vectors are normalized into a uniform scale.

Ivana Arsic [40] have consider the problem of automatic extraction of the geometric lip feature for the purpose of multi-modal speaker identification. The use of visual information from the mouth region can be great importance for improving the speaker identification system performance in noisy conditions. They proposed method
for automated lip feature extraction that utilizes color space transformation and a fuzzy based c-means clustering technique. Using the obtained visual cues closed-set audio-visual speaker identification experiments are performed on the CUAVE database and method resulted to promising results.

Leon J M Rothkrantz [41] presented a way of processing the video signal for lipreading applications and a post processing data transformation that can be used alongside it to improve the audio-visual speech recognition results. They represented Lip Geometry Estimation (LES) method and was compared with other geometry and image intensity based technique typically deployed for the said task. It can be applied at post processing stage to any other feature extraction technique. They have shown that at what extent different ways of processing the video signal are equivalent under appropriate transformations.

Gang Li [42] has discussed that, a speech synthesis is designed for automatically recognizing visual speech generated by the speech impaired and present a new communication approach for the speech impaired people. In order to acquire more parameters of lip contours, a model was framed, which can extract the degree of pouting form it. At the same time, the differential coefficients of some parameters are calculated to describe dynamic characteristics of the lip contours. Movement detection and morphological processing are used to extract mouth area and parameters of lip contours from the image sequences.

Maycel Isaac Faraj [43] described motion based feature extraction technique for speaker identification using orientation, estimation in 2D manifolds. The motion estimation was done by computing the components of the structure tensors from which normal flows are extracted. By projecting the 3D spatiotemporal data to 2D planes they have obtained projection coefficients which they used to evaluate 3-D orientations of brightness patterns in TV like image sequences. This corresponds to the solutions of simple matrix Eigen value problem in 2D, affording increased computational efficiency. An implementation based on joint lip movements and speech was presented along with experiments which confirm the theory, exhibiting a recognition rate of 98% on the publicly available XM2VTS database.

Alin G Chitu [44] extended their earlier work [41] and discussed the techniques used for extraction of static visual features result in equivalent features or at least the most informative feature that exhibits the property. They have analyzed the importance of motion detection for speech recognition. For this considered Lip
Geometry estimation for static feature extraction. This method combines an appearance based approach with statistical based approach for extracting the shape of mouth. Further they introduce a method based on approach that captures relevant motion information with respect to speech recognition by performing optical flow analysis on the contour of speaker mouth. The evaluation of these methods was done under different noise conditions. They have shown that audio-video recognition based on the true motion features that are obtained by performing optical flow analysis outperforms the other settings in low signal to noise ratio (SNR) conditions.

Abu Sayeed Md [45] describes a fully automated technique of detecting lip contour from static face image. Face detection is performed first on the input image using a variation of the AdaBoost classifier trained with Haar like features extracted from face. A trained classifier is applied over this extracted face region for isolating the mouth section. The detection of lip contour is performed from this isolate mouth region using the level set method of image segmentation. The variation formulation was proposed by Li et.al (CVPR 2005) and has been applied for the problem. This completely eliminates the need of re-initialization procedure. The proposed method has been tested on three different face databases that contain images of both natural faces as well as facial expressions and maximum successful lip contour detection rate of 91.08% was achieved.

Takeshi Saitoh [46], Analysis of efficient lipreading method for various languages where they focused on limited set of words from English, Japanese, Nepalese, Chinese, Mongolian. The words in English and their translated words in above listed languages were considered for the experiment. They have presented new feature called trajectory feature, where trajectory feature is a time change of n feature expressed as a n-dimensional trajectory of lip motion of target words. The trajectory was generated by plotting points in n-dimensional space. For recognition method they presented dynamic programming and matching of word. The databases of words have been collected as 10 samples of each word from one subject per language using digital video camera. The utterance scenes were recorded with three students those known to Japanese, Mongolian, Nepalese with image size of 640x480 pixel and frame rate of 30 frame per second.

Meng Li [47] in the work on A Novel Motion Based Lip Feature Extraction for Lip-reading has addressed to the issue of how to extract the visual features, which greatly impact on the lip-reading recognition accuracy and efficiency. The method
does not utilize all pixels in each image, but just some landmarks with the obvious lip movement. Since the lip movement for each utterance is different, the positions of landmarks are various in each image. Further, the number of these landmarks may not be constant. When the landmarks are chosen, the corresponding movement tracks will also captured, whereby the feature is built to represent the lip movements.

More comprehensive reviews of psychological aspects of human speechreading may be found in [48, 49, 50, 51]. It is clear however, that even normal hearing humans are able to make good use of visual speech cues and that is particularly important under noisy conditions. The McGurk effect shows that auditory and visual speech perception are tightly bound and appear to be so from an early age.

3. **Speech Production**

A schematic diagram of the human vocal mechanism is shown in Figure 2.1 (a) and Figure 2.1(b) presents principal features of vocal tract. The air enters the lung via the normal breathing mechanism. As air is expelled from the lungs through trachea (or windpipe), the tensed vocal cords within the larynx are caused to vibrate (in the mode of relaxation oscillator) by the air flow. The vocal cords are two muscular folds that are usually apart for breathing, but can be brought close together and then vibrate in airstream from lungs. This vibration is controlled by the tension of the chords and modulates the airstream. This process of phonation, and the sounds produced are voiced. Sounds made using an unrestricted, un-modulated airstream are unvoiced. Depending on the position of the various articulators that is jaw, tongue, velum, lips, mouth different sounds are produced [52].

![Figure 2.1](image)

**Figure 2.1** (a) Schematic view of human vocal mechanism, (b) Principal features of vocal tract [53]
Above the larynx is the vocal tract, the first stage of which the pharynx (back of throat) which can be tightened to change speech sounds, but it is not often used in English. For the pharynx the airflow may be redirected either into nose and mouth, or just the mouth by closing velum (soft palate). Sounds made with the velum open are nasal and with velum closed oral. The shape and configuration of the vocal tract further filter the speech sound. The sounds produced can be classified according to the place and manner of their articulation. The manner of articulation describes the degree of occlusion used to create sound. For example, a complete closure and hold of the articulators halts the airstream and is called a stop, example the first and last sounds in ‘pop’. A fricative occurs when the articulators are brought close enough together to cause a turbulent airflow, ex ‘shoo’ and an approximant is when articulators are close, but not enough to cause a fricative, e.g. ‘we’. Finer classification of the manner of articulation can be made and are described in [53, 54].

The place of articulation described which articulators are used, and is classified as one of

- **Bilabial** between both lips, for example ‘pie’
- **Labiodentals** between lower lip and upper teeth, for example ‘fie’
- **Dental** between tongue tip or blade and upper front teeth, for example ‘thigh’
- **Alveolar** between the tongue tip or blade and alveolar ridge, for example ‘tie’
- **Retroflex** tongue tip and back of the alveolar ridge, this is not used in English
- **Palato-alveolar** tongue blade and back of the alveolar ridge, for example ‘shy’
- **Palatal** between the front of the tongue and hard plate, for example ‘you’
- **Velar** back of the tongue and the soft palate, for example, the end of ‘hang’

If there is no contact between articulators the sound is a vocoid (as opposed to contoid) and there is no identifiable place of articulation. These are classified, using tongue position in the cardinal vowel space, as front, center or back, and low, mid or high. Additionally, lip shape may be classified as rounded or spread.

### 4. Speech sounds and features

The number of linguistically distinct speech sounds (phonemes) in a languages is often matter if judgment and is not invariant to different linguistics.

#### 4.1 Phonemes and Visemes

When two different speech sounds can be used to change the meaning of the word they are different phonemes. A phoneme is an abstract unit of speech sound that is useful for transcribing speech in an unambiguous manner. Because there are more
phonemes than letters in mist alphabets, the International Phonetic Association (IPA) defines phonetic alphabets for transcriptions that are common across all languages and for devnagari scripts the IPA symbols as a place of articulation are described with reference to articulator and vowels, consonants and their arrangement is shown in Table 2.1 (A) and 2.1 (B) [55].

<table>
<thead>
<tr>
<th>Articulator</th>
<th>Independent form</th>
<th>Romanized</th>
<th>As diacritic with</th>
<th>Independent form</th>
<th>Romanized</th>
<th>As diacritic with</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Guttural)</td>
<td>OrderId</td>
<td></td>
<td></td>
<td>OrderId</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Palatal)</td>
<td>OrderId</td>
<td></td>
<td></td>
<td>OrderId</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Labial)</td>
<td>OrderId</td>
<td></td>
<td></td>
<td>OrderId</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Retroflex)</td>
<td>OrderId</td>
<td></td>
<td></td>
<td>OrderId</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Dental)</td>
<td>OrderId</td>
<td></td>
<td></td>
<td>OrderId</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Caudal)</td>
<td>OrderId</td>
<td></td>
<td></td>
<td>OrderId</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Labial-Guttural)</td>
<td>OrderId</td>
<td></td>
<td></td>
<td>OrderId</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.1 (A) Articulator and corresponding vowels and their arrangement [55]

<table>
<thead>
<tr>
<th>Articulator</th>
<th>Independent form</th>
<th>Romanized</th>
<th>As diacritic with</th>
<th>Independent form</th>
<th>Romanized</th>
<th>As diacritic with</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Stop)</td>
<td>OrderId</td>
<td></td>
<td></td>
<td>OrderId</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Nasal)</td>
<td>OrderId</td>
<td></td>
<td></td>
<td>OrderId</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Approximant)</td>
<td>OrderId</td>
<td></td>
<td></td>
<td>OrderId</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Fricative)</td>
<td>OrderId</td>
<td></td>
<td></td>
<td>OrderId</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.1 (B) Articulator and corresponding consonants and their arrangement [55]

4.2 The Vowels

The vowels are produced by exciting an essentially fixed vocal tract shape with quasi-periodic pulses of air caused by the vibration of the vocal cords. The way in which the cross-sectional area varies along the vocal tract determines the resonance frequencies of the tract (the formants) are thereby the sound that is produced. The vowel sound produced is determined primarily by the position of tongue, but the
position of jaw, lips and to the small extent, the velum, also influences the resulting sound. The vowels are generally long in duration (as compared to consonant sound) and are spectrally well defined. As such they are usually easily and reliably recognized and therefore contribute significantly to our ability to recognize speech, both by human being and by machine. A wide range of variability can be seen in measured formant frequencies for a given vowel sound and also there is overlap between the format frequencies for different vowel sound by different talkers.

4.3 Semivowels
The group of sounds consisting of /w/, /l/, /r/ and /y/ is quite difficult to characterize. These sounds are called as semivowels because of their vowel like nature. They are generally characterized by a gliding transition in vocal tract area function between adjacent phonemes. Thus the acoustic characteristics of these sounds are strongly influenced by the context in which they occur.

4.4 Nasal consonants
The nasal consonants /m/, /n/ and /ŋ/ are produced with glottal excitation and the vocal tract totally constricted at some point along the oral passageway. The velum is lowered so the air flows through the nasal tract, with sound being radiated at the nostrils. The oral cavity although constricted toward the front, is still acoustically coupled with pharynx. Thus the mouth serves as a resonant cavity at the traps acoustic energy at certain natural frequencies. For /m/ the constriction of lips; for /n/ the constriction is just behind the teeth; and for /ŋ/ the constriction is just forward of the velum itself.

4.5 Voiced and unvoiced stops
The voiced stop consonants /b/, /d/ and /g/ are transient, noncontinuant sounds produced by building up pressure behind a total constriction somewhere in the oral tract and then suddenly releasing the pressure. For /b/ the constriction is at the lips; for /d/ the constriction is at the back of the teeth; and for /g/ it is near the velum. During the period when there is a total constriction in the tract, no sound is radiated from the lips. However there is a often a small amount of low-frequency energy radiated through the walls of throat (sometimes called as voice bar). This occurs when the vocal cords are able to vibrate even though the vocal tract is closed at some point. Since the stop sounds are dynamic in nature, their properties are highly influenced by the vowel that follows the stop consonants.
5 Visual features

Vision has been shown to be important in human speech perception, and then the question often comes in mind that what are the visual cues that best provide speech information?

Benoit [05] compared speech intelligibility under conditions of audio-only, audio and lips, audio and face for natural and synthetic image displays. The lips gave improved intelligibility but this was further improved by the addition of the rest of the face. The similar results were presented by McGrath [56], who found that being able to see the teeth improved performance compared to just the lips. Jackson and Montgomery [57, 58] used statistical analysis of physical lip measurements in a vowel task. They found the strongest dimensions were horizontal and vertical lip opening. There is some evidence that is the dynamic motion of the visible articulators that is most important. Rosenblum [59, 60] found a kinematic point light display of lower face could be used to induce McGurk effects. The only visual features were dots placed on the talkers face. However, neuropsychological studies of brain damaged patients by Cambell [61, 62, 63] finds conflicting examples. A patient unable to recognize static faces or read facial expression was found to have normal speechreading ability and responded to McGurk illusions. Another patient, that could see motion but not form, was unable to speechread. She concluded that both visual form and movement are required for speech reading [64].

Munhall and Bateson [65, 66, 67] describe a combined optical point tracking and electromyography analysis of facial dynamics during speech. They found visual speech information was distributed over the entire face and not just the lips. They also found that even in high noise condition the listener did not look at the lips of talker more than 50% of the time. This suggest that, human are able to perceive visual speech using the low spatial resolution off-fovea parts of the retina. Experiments by Jorden [68] have shown that speechreading is robust over large variations in image size.

6 Summary

In summary, while most visible speech information is, unsurprisingly, seen around the lips and it is also distributed over the entire face. Whenever the component is removed for example face or teeth, the intelligibility falls. Visual speech analysis should take care when discarding some of the many possible facial features which are available. The methods discussed in review are primarily based on Contour Coding,
Dynamic Programming, Multi-State Time Delayed Neural Network, Active Shape Model, Hidden Markov Model, Active Contour Model, Discrete Cosine Transform, and Principal Component Analysis etc. The Principal Component Analysis is widely chosen for dimension reduction and act as basis for building shape models.
7 References
[02] J Bulwer. Philocoupus, or the deaf and dumbe mans friend, Humphrey and Moseley, 1648
[05] C Benoit, T Guiard-Marigny, B Le Goff and A Adjoudani, Which components of face do humans and machines best speechread?, In Stork and Hennecke [63], pp 315-328
[16] Christoph Bregler, Stephen M Omohundro, Surface learning with applications to lipreading, Appeared in J.D Tesauro, G., and Alspector J, Editors, Advances in Neural Information processing systems, 6 San Fransisco, CA, Morhgan Kaufmann Publisher


[40] Ivana Arsic, Roger Vilagut and Jean-Philippe Thiran, Automatic extraction of geometric lip features with application to multimodal speaker identification, IEEE International Conference on Multimedia and Expo, page 161-164, 2006

[41] Leon J M Rothkrantz, Jacek C Wojdel, Pascal Wiggers, Comparison between different feature extraction techniques in lipreading applications, 11-th International Conference SPEECH AND COMPUTER (SPECOM 2006),St. Petersburg, 25-29 June 2006.


[51] Q Summerfield, Lipreading and audio-visual speech perception, Phil Trans, R.Soc. Lond B 335:71-78, 1992


[60] L D Rosenblum and H m Saldana, Time-varying information for visual speech perception, pages 61-81


[65] K G Munhall and E Vatikiotis-Bateson, The moving face during speech communication, In Campbell et.al [64], pages 123-139.

[66] E Vatikiotis-Bateson, K G Munhall and M Hirayama, The dynamics of audiovisual behavior in speech, In Storke and Hennecke [63], pages 221-232
