Implication and Utilization of various Lip Reading Techniques

Saiqa Khan
Professor, Department of Computer Engineering, M.H. Saboo Siddik College of Engineering, Mumbai, India

Hamza Azmi
Student, Department of Computer Engineering, M.H. Saboo Siddik College of Engineering, Mumbai, India

Ajay Nair
Student, Department of Computer Engineering, M.H. Saboo Siddik College of Engineering, Mumbai, India

Hamza Mirza
Student, Department of Computer Engineering, M.H. Saboo Siddik College of Engineering, Mumbai, India

ABSTRACT
Ears are one of the most important sensory organs of the human body. It is one of our primary sensors along with our eyes, nose, skin etc. However, while many people may have perfectly functional ears some suffer from partial or total hearing loss or hearing impairment. In order to tackle this, many suggest the use of hearing aids, sign language, cochlear implants. But of all the different ways, one of the best way to beat this inability is to train oneself to lip read. In simple words, lip reading is the interpretation of the movements of not only the lips but also the face and tongue. While many people are excellent at lip reading, some find it a challenge too difficult to overcome. In regard to this, one cannot deny that whenever some human falls short of something, a man-made technology can make up for the shortcoming to a certain extent. As a result, over the years a lot of time and effort has been invested in studying various lip reading algorithms and speech automations. Through this research paper, the aim is to put forth a study on various lip-reading algorithms, their approaches, their innovations and the central ideas behind their proposed studies.

Keywords
Hearing loss, Hearing impairment, Lip Reading, Speech Automation.

1. INTRODUCTION
People who are unable to hear and perceive sound signals are said to suffer from hearing loss. As seen in Figure 1, the human ear can be divided into 3 sections namely outer ear, middle ear and inner ear.

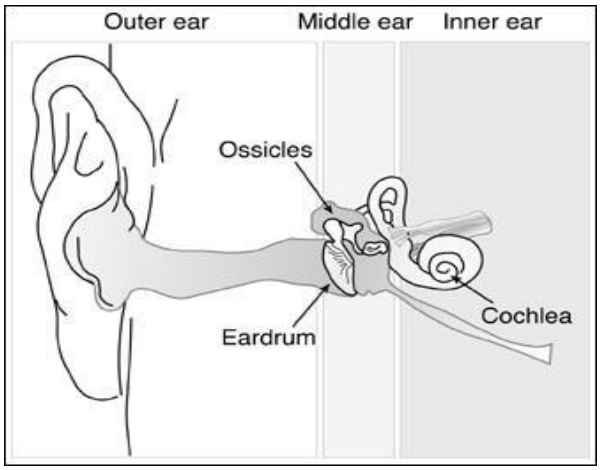

Figure 1: The Human Ear

Depending on whether the hearing impairment is due to the outer, middle, inner ear or all three; we can classify the hearing loss as conductive, sensor neural, central and functional. Table I maps the different types of hearing losses along with their causes.

Table I: Different Types of Hearing Loss

Hearing Loss	Part of the Ear	Causes
Conductive	Outer, Middle	Allergies, Fluid in middle ear, Perforated eardrum
Sensor neural	Middle, Inner	Tumors, Age factor, Loud noise
Central	Middle, Inner	Damaged Auditory nerve, Trauma
Functional	-	Depression

The hearing impairment can also be classified as mild impairment (minimum sound that can be heard is around 25-30dB), moderate impairment (minimum sound that can be heard is around 40-70dB) and severe impairment (minimum sound that can be heard is around 70-95dB). Depending on the level of impairment an impaired person can opt for a suitable way to tackle his/her impairment.

While there may be many options available to a hearing-impaired person today, like for example hearing aids, assistive devices, cochlear implants etc., it is important to note that a lot of drawbacks come along with it. For arguments sake, the hearing aid technology fails to help or assist people in an environment where the noise and disturbances swamps out the voices of the individuals. On researching the various ways to effectively deal with hearing loss, one of the techniques which caught our attention was the lip-reading technique. Lip reading is the technique where the shape of the lip maybe analyzed in order to predict accurately the word that is being said. Here we primarily analyze the images of the lips of the speaker along with the movements of the lips in order to understand the words of the speaker. A central part of any lip-reading algorithm is to extract the images of the lips. After this the processing of these images are carried out with the aim of yielding correct outcomes.

Various authors have come up with various ingenious ways to process the images and get the desired results. Through this paper, we aim to study and understand a few of the different previous studies. The next section speaks in brief about the various previous studies.

2. LIP READING

When it comes to lipreading, the appearance or the shape the pic makes while talking is key to understanding the letters or the words that are being spoke. Figure 2 gives an idea of the shapes that the lip makes with respect to certain letters or words. We now move forward with the paper, by speaking in detail about the various previous studies and researches.

Figure 2: Appearance of lip during various sound.

The paper [1] takes into consideration the importance of automatic speech recognition when the environment is nosier than usual. Through this paper, the authors propose the use of RASTA, which is a type of inter-filtering method that helps in reducing stationary and convolutional noise that maybe present, along with Image Transform Lip Reading Algorithm (ITLR) with the aim of enhance the overall performance of automatic speech recognition systems in general.

In a nutshell, the authors propose the integration of inter-frame filtering with the lip-reading algorithm. In order to achieve this, the authors propose the bringing together of the two phases namely PRE-I (pre-integration) and POST-I (post-integration). As the name suggests, in the PRE-I phase the RASTA and inter-frame filtering is performed prior to the image transform whereas in POST-I phase, inter-frame filtering is done after the image transform process.

On observing the experimental results, one can easily pass a verdict in favor of the proposed technique.

The paper [2] works on active appearance model. It works on the idea of getting support from cavity features, like appearance of teeth and tongue, when the movement of the lip can't be precisely tracked or even if after doing so it doesn't contribute much help. For this the paper proposes the idea wherein it uses DCT or DWT as the initial step in the active appearance model.

In lip reading techniques, lip detection or segmentation is the most basic step and a challenging one too. Most researchers take the help of the red coloration of the lips to segment the lips. But this may not provide us with accurate results in case of poor choice of background lights, red blobs in the speaker's clothing or any other of such unforeseen factors. To counter this problem, the author of the paper made use of Adaboost algorithm for face and mouth detection. Adaboost classifier cascades Haar like features, hence it is fast and accurate for face detection [3]

Transformations are often accompanied with a dimension reduction phase like Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA) requiring finite number of computation for training as well as testing. PCA focuses on reducing the data size by diminishing the difference between the actual and the recovered data. LDA aims for better segmentation and keeping the discriminating features intact.

In this paper [4], S.L.Wang et al. extracted the lip features by taking in account lip contours. The features extracted by lip contour extraction were:

i) Geometric Features:
This gives the details about the width and height of the lips.

ii) Shape Descriptors:
Active Shape Model (ASM) is used for the extraction of this feature. Contour point coordinates for any shape x can be approximated as

$$x = x + Pb \qquad (1)$$

where
x = mean shape,
P = matrix of eigen vectors of the covariance matrix
b = weight vector for each eigen vector.

In the proposed approach, Wang made use of the first three weight values as they are sufficient for modeling shape variation

iii) Inner mouth features: teeth and mouth opening
From the technique in [7], teeth region area is calculated and normalized against the entire mouth, which is then enclosed by the outer lip contour. Inner mouth area is yet another parameter that is taken into account, this is obtained by analyzing luminance of the teeth and teeth information along the central axis of the mouth.

Spline curve has been used while performing the experiment, to produce static and dynamic information for each visual feature. From all the training samples, spline coefficients have been estimated in order to avoid over fitting.

Experimental results displayed that, when only limited data is available, accurate results are obtained for speaker dependent and speaker independent tests.

In the previous study [5], Breger et al. performed experiments with the intention of improving automated speech perception's recognition performance by taking the help of lip reading. Which can also be called as 'Speech reading'.

For carrying out the experiment, Breger et al. applied two techniques for preprocessing: Histogram normalized grey-value coding, or 2 dimensional Fourier transformation. In either of the cases they only considered an area of interest which encompasses the lips and then applied low pass filter on them.

It presented an extension of the existing Multi-State Time Delay Neural Network architecture (MS-TDNN) [[6] H. Hild and A. Waibel. Connected Letter Recognition with a Multi-

State Time Delay Neural Network. Neural Information Processing Systritis (NIPS 5)] for taking into consideration modalities, acoustic and visual sensor input. It projects how a system with amalgamated acoustic and visual information is better than system that makes use of acoustic only.

In this paper [6], the system proposed by Ralph Krike et al. operates under the influence of active near infrared illumination. It makes use of a technique which makes use of local binary patterns, to model lip motions with hidden Markov models. It evaluates accuracy with TUNIR database.

Hidden Markov Models (HMMs) was used to classify lip analysis feature's sequences. The HMMs are trained using the Baum-Welch re-estimation procedure. The classification of an unknown feature sequence is performed with the Viterbi algorithm.

They evaluated the system with 9 different speakers uttering digits zero to nine from the TUNIR database [10]. They used three sequences for training and the remaining sequences were used as test. This procedure was repeated 4 times, using every sequence as test set. Performance of the classifier was reported as an average of the four test.

Finally, Ralph Kricke et al. concluded that local binary

pattern seems to be best choice for the examined classification tasks in terms of classification accuracy for still images by including spatial information. If rotation invariance was taken into consideration it would have an undesirable impact on the operator's performance. It was even seen that, the use of HMM for each digit performs well although little training data available to them.

3. COMPARISON

The purpose of the literature survey in this paper was to identify various studies, models and papers in our proposed research area i.e. Lip- Reading, in an attempt to appreciate, make use of as well bridge a missing gap, if any, between different researches. Section II provides a general study on the various previous studies by giving an idea about how the authors thought about moving forward with the lip-reading challenge.

In this section, we compare a few of the above-mentioned techniques. We consider parameters such as Feature Extraction, Number of Frames Selected, Media used, Language used for pattern matching and the Number of test subjects.

Following is the table which gives a comparative study on the different researches (Table II).

Table II: Comparison of Various Techniques

	Feature Extraction	**No. of frames selected**	**Media Used**	**Language Used**	**No. of Test Subjects**
[1]	Lip folding, Principal component analysis	30	Video	Korean	70
[2]	2D DWT with three decomposition levels, PCA, LDA and LSDA	16	Video	English	7
[4]	Geometric features are normalized, Active Shape Model	25 or 30	Video	English	8
[5]	Grey-Value coding, 2D-FFT	30	Video	German	2
[6]	Hidden Markov Models, Baum-Welch re-estimation procedure, Viterbi algorithm	-	Image	English	90

4. CONCLUSION

There are number of techniques available for lip reading and speech recognition. Each Technique has a downside as well as some up points. While choosing the best technique, we have to take into consideration various factors such as can we give initial training to the machine or the test will be completely new. We also need to check our lighting condition for proper face and mouth detection.

With technology moving towards its zenith, many implementations come to light from time to time. A lot of ideas are proposed prior to the implementations. Through our research paper, we put together a lot of these ideas proposed by people all around, with the aim of motivating new innovations in the field of hearing disorder.

5. REFERENCES

[1] Jinyoung Kim, Suengho Choi, Seongmo Park "Performance Analysis of Automatic Lip Reading Based on Inter-Frame Filtering".

[2] Lip Reading Using DWT and LSDA Sunil Sudam Morade and Suprava Patnaik Professor, Advance Computing Conference (IACC), 2014 IEEE International 27 March 2014.

[3] P. Viola, M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple features", IEEE Int. Conference, 511-517, 2001.].

[4] Automatic Lipreading With Limited Training Data , The 18th International Conference on Pattern Recognition (ICPR'06), S.L.Wang, W.H.Lau, A. W. C. Liew and S.H.Leun.

[5] Bregler, C., Hild, H., Manke, S., and Waible, A 1993. Improving connected letter recognition by lip reading. In proc. IEEE Int. Conf. on ASSP, pp. 557-560.

[6] Local Binary Patterns For Lip Motion Analysis, Ralph Kricke, Thorsten Gernoth, Rolf-Rainer Grigat, 978-1-4244-1764-3/08/$25.00 ©2008 IEEE.

[7] K.L. Sum, W.H. Lau, S.H. Leung, A.W.C Liew, K.W. Tse, "A new optimization procedure for extracting the point-based lip contour using active shape model", *Proc. Of ICASSP'01*, Salt Lake City, Vol.3, pp.1485-1488, 2001.

[8] Lip Reading www.lipreading.org

[9] Lip Reading https://en.wikipedia.org/wiki/Lip_reading

[10] S. Zhao, R. Kricke, and R.-R. Grigat, "Tunir: A multimodal database for person authentication under near infrared illumination," in *Proceedings of 6th WSEAS International Conference on Signal Processing, Robotics and Automation (ISPRA 2007)*, Corfu, Greece, February 2007.