DOI: http://dx.doi.org/10.26483/ijarcs.v14i3.6970

Volume 14, No. 3, May-June 2023

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

SEMG APPROACH FOR SPEECH RECOGNITION

Siddesh Shisode, Bhavesh Mhatre, Jeet Sikligar, Sushil Vishwakarma,
Supriya Tupe, Sheetal Jagtap and Milind Nemade
Department of Electronics KJSIT
Mumbai, India

Abstract- Speech is the most familiar and habitual way of communication used by most of us. Due to speech disabilities, many people find it difficult to properly voice their views and thus are at a disadvantage. The research tackles the issue of lack of speech from a speech impaired user by recognizing it with the use of ML models such as Gaussian Mixture Model - GMM and Convolutional Neural Network - CNN. With properly recorded and cleaned muscle activity from the facial muscles it is possible to predict the words being uttered/whispered with a certain accuracy. The intended system will additionally also have a visual aid system which can provide better accuracy when used together with the facial muscle activity-based system. Neuromuscular signals from the speech articulating muscles are recorded using Surface Electro Myo Graphy (SEMG) sensors, which will be used to train the machine learning models. In this paper we have demonstrated various signals synthesized through the ElectroMyography system and how they can be classified using machine learning models such as Gaussian Mixture Model and Convolutional Neural Network for the visual-based lip-reading system.

Keywords- EMG, Speech Recognition, Silent Speech, GMM and CNN

I. INTRODUCTION

The research tackles the issue of lack of speech from a speech impaired user by recognizing it with the use of ML models. With properly recorded and cleaned muscle activity from the facial muscles it is possible to predict the words being uttered/whispered with a certain accuracy. The neuro-muscular activity will be recorded from surface electrodes; it will be a non-invasive method, meaning it does not involve the electrodes being inserted inside the body. Since the method being used is a non-invasive measure to record the muscle activity, it might contain other noises in the signal, these can be removed easily but to ensure a higher accuracy, a visual method to track what the user is intending to speak can be used. The visual based system will record the lip-movement of the user with the help of a camera and output a list of probable words the user is intending to speak. By pairing the EMG based system with the visual system, it is possible to predict the correct word. The intended system would be interfacing of both, the EMG and Visual based system which will let us, in theory, be able to predict the syllables and consonants thus effectively predicting the actual word [6]. It can be especially useful for people who have had Laryngectomy, where in the voice box of the person is removed. In such a situation, the person is not able to produce any sound while speaking but is able to move his lips and other sub vocal muscles[12]. With the help of such a system, people with speech disabilities will be able to communicate easily with other people who do not have adept knowledge of sign-language. In this paper we have proposed a system which will take the EMG signal as input from the user and apply Gaussian Mixture Model to classify the given EMG signals to the corresponding spoken word.

The further part of the paper is divided into sections which explain the previous work done by researchers, the EMG signal synthesis and its classification process, Visual speech feature extraction and model training and finally demonstrating the EMG system's signal output and our inferences from it.

A. Abbreviations

Convolutional Neural Network - CNN, Gaussian Mixture Model - GMM, Motion History Index - MHI, Principal Component Analysis - PCA, Audio Speech Recognition - ASR, Multiscale Spatial Analysis - MSA, 2-Dimensional Linear Discriminant Analysis - 2DLDA, Bidirectional Long Short-Term Memory – BLTSM

II. RELATED WORK

There have been numerous speech-reading techniques reported in various literature papers. The early studies date as long back as 1980s, where Sugie and Tsunda worked in this domain. Sugie and Tsunda used 3 electrodes to capture the EMG signals of face muscles. They used the EMG signal to recognize Japanese vowels[1].

Morse et al designed a system to classify among 10 English words. This work was the extension of their previous research where they had designed a system to discriminate between 2 words using neck and head muscles[13], [14].

Matthias Janke and Lorenz Diener, recorded the EMG signals for audible speech and converted the EMG signal to speech using GMM. Such a method is not limited to a given phone-set, vocabulary or language[6].

Further, Matthias Janke with M. Wand continued the research in this domain. Where in they denoted the EMG signals as static source and the actual speech as target output vectors and trained the GMM, which would relate the target and source vectors using the joint probability density[7].

Wai Chee Yau et al discuss a method of using SEMG signal in conjunction with a visual system. Wai Chee Yau et al suggested that it is easier to decode vowels through the EMG signals, and through visual system, consonants can be decoded, thereby they were able to predict the phoneme the

subject was trying to utter with a accurate mean classification rate of 84.7% for certain English visemes[19]. Longbin Lu et al compared different lipreading techniques for silent speech interfaces. Most techniques follow the same process of converting the image frames to gray-scale images and then extracting the features from these images using transformations like PCA, MSA. But Longbin Lu et al suggested that using model based methods for feature extraction methods are more useful since their performance is not hindered because of rotation or translation of the lip contours. They proposed a Homeomorphic Manifold Analysis (HMA), which initially maps a non linear mapping of the inputs with unified mapping, In the further steps using 2DLDA, the features can be extracted from the mapping matrix and then the speech can be classified by applying Support Vector Machine on these extracted features[10].
Based on the papers referred, it is clear that although the SEMG signals can be used to identify the spoken speech, a supporting system based on the visual system will help increase the accuracy of the system.

III. EMG SIGNAL SYNTHESIS

EMG stands for electro-myography, where the neuro-muscular activity of muscles can be monitored using EMG electrodes and sensors.

A. EMG Electrodes When triggered, the neurons in the human body emit a signal of about 10 mV. In this approach, non-invasive bipolar electrodes will be used to record these signals. SENIAM (Surface ElectroMyoGraphy for the Non-Invasive Assessment of Muscles) guidelines were used to determine the characteristics of the electrodes utilised in the system. The analysed muscle being a small muscle, these rules state that the bipolar electrodes should be positioned closer to these muscles and they should be compact in size. The electrodes must be placed on a dry, smooth surface. The adhesive qualities of the surface electrodes are hampered by skin hairs, perspiration, and wrinkles. We employed numerous channels using a set of three electrodes in our study. The reference electrode can be placed on the forehead, elbow or any skin surface having no muscle interference [1], [2], [4].
   - Zygomaticus Major
   - Orbicularis Oris
   - Levator labii sup
   - Depressor Labii inf.
   - Depressor angulioris

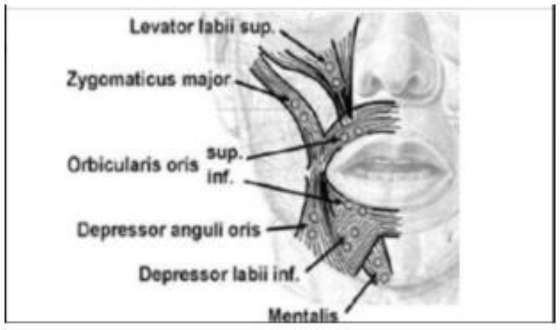

Fig. 1. Topological location of muscles

Fig. 2. Electrode Placement

Fig. 3. Electrode Placement

B. Signals Acquired The following are the signal graphs plotted as received from the muscle sensors. The words or utterances for which thesignals are being retrieved are medical or health related terms, that a patient may use - Emergency, Heart, Water, etc.

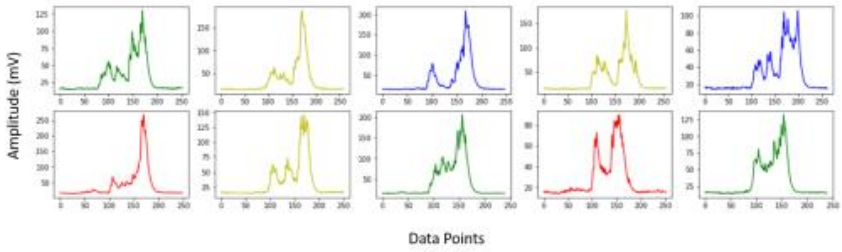

Fig. 4. Signal acquired for 'Medicine'

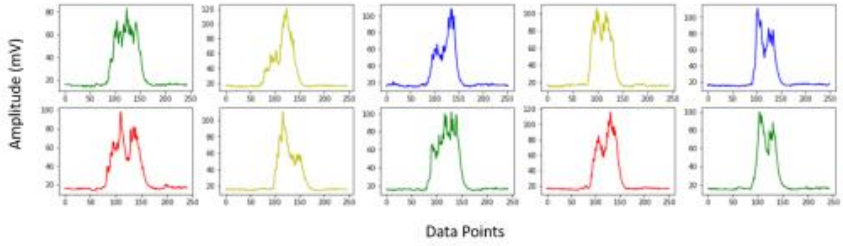

Fig. 5. Signal acquired for 'Bring'

C. Algorithm - Gaussian Mixture Model(GMM) The signals as acquired from the sensors are cleaned, organized, trimmed and passed to a GMM algorithm. A Gaussian Mixture is a function compromising of many Gaussians, each of which is distinguished by a unique identifier. {kϵ{1, 2, ..., K}} where the number of clusters is defined by K. The following parameters define each Gaussian k in the mixture:

• A centre which is defined by its mean μ.
• The width, which corresponds to its covariance P. In a multivariate situation, this corresponds to the dimensions of an ellipsoid.
• A π probability mix that determines the magnitude of the Gaussian function.

Numerous datasets may be represented by the Gaussian Distribution (Univariate or Multivariate). It attempts to represent the dataset as a combination of many Gaussian Distributions. This is the model's central concept. A Gaussian Distribution's probability density function in one dimension is given by the following formula.

$$G(X|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad (1)$$

where μ and σ 2 describe the mean and standard deviation for the specified distribution. The probability density function of a Multivariate Gaussian Distribution (say d-variate) is given by the following equation :

$$G(X|\mu,\sigma) = \frac{1}{\sqrt{(2\pi)|\sum|}} e^{-(\frac{1}{2}(X-\mu)^T \sum^{-1}(X-\mu))} \quad (2)$$

A corpus of utterances in which the EMG signal and auditory signal have been concurrently captured is required for training.Source data for the GMM training consists multiple utterances of the same word stacked up as matrices which are fed to the model. Note that the 0th Mel coefficient is not used since it indicates the acoustic signal's strength, which is difficult to quantify using EMG [17]. The model then creates a GMM model for each individual word. This model file calculates the probability of the input signal belonging to that particular class.

IV. EXTRACTING VISUAL SPEECH FEATURES

For lip-reading the visemes, phonemes from the subject, it is necessary to have a streamlined process. The visual system will record the lip-activity for each word utterance, and the video feed needs to be converted into individual frames to be able to analyse them for the machine learning models. For the classification of different words/visemes from these frames, a CNN model can be used [20], and the process to do so is discussed below.

A. Algorithm - Convolution Neural Network (CNN) + Long Term Short Memory (LSTM)

Neural networks such as the Convolutional Neural Network are particularly well-suited for image processing tasks, such as object recognition and classification. Each word utterance maps to a fixed way in which the lips move. A CNN model can learn these patterns and classify them accordingly. In

order to successfully classify them, following procedure needs to be followed:

1) Dataset requirement: In order to train a CNN model, a large dataset containing ordered and labeled frames for each word utterance is required. The dataset will contain four to five frames for each utterance where the lip contour movement is clearly visible. We have used MIRACL-VC1, which is a lip-reading dataset [15] which contains sequence of images for a sequence of words for fifteen people.

Fig. 6. Frames showing different utterances

2) Preprocessing: The individual frames need to be standardized by having a uniform size and cropping the image such that lip contours are at the center of the frame. This is done using face detection in image module from Python. Then the image is converted into machine readable image code values [3], [9]. Additionally, using a motion based differentiation technique called as spatial-temporal-templates(STT), images can be generated from the video feeds. STT segmentation is a technique which shows area of high movement in a grey-scaleimage . It does so using an approach known as image difference [19].

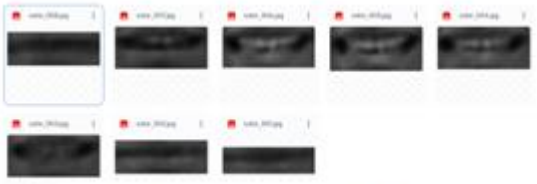

Fig. 7. Cropped and greyscaled images

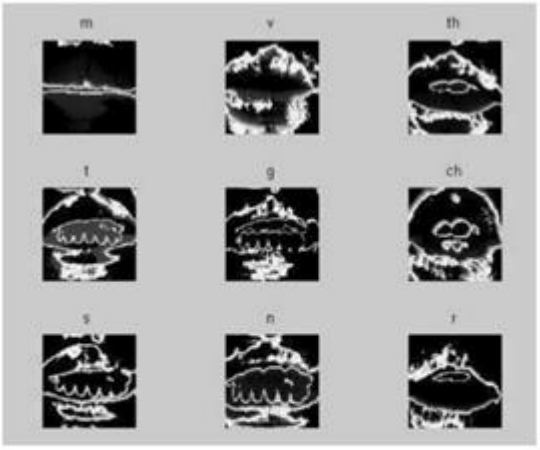

Fig. 8. Example of STT for different alphabet utterances

$$STT_i = max \bigcup_{i=1}^{i=N-1} B_i(x,y) * i \qquad (3)$$

Here N is used to represent the number of frames in the video-file, and Bi(x,y) represents the binary version of the ith frame.

3) Training & Validation: The next step in the process is to feed the standardized labeled data to the CNN model for training. Based on the accuracy of the model using validation data, parameters of the CNN model can be tuned. These parameters can be number of pooling and convoluting layers in the network, number of neurons in each layer, etc. An improved method to enhance the accuracy levels would be to use 3D-2D-CNN-LSTM-based model which utilizes height, width and time or sequential parameters [11]. Since uttering a word is a sequential process, including the time parameter makes sense. The 3D-2D-CNN-BLTSTM model has 2 CNN models, followed by a pooling layer of BLSTM, which helps the model to temporarily hold the information about the previous frame(s). Hence the name, Bidirectional Long ShortTerm Memory (BLSTM). Alternatively, we may also use Long Short-period storing cyclic neural network, which works on the principle of feed-forward neural systems [16] along a DNN to identify the features or a CNN + LSTM architecture. Similarly, other variations of feature extractions can be used along with CNN such as Zoning, LDA [8], [18], KarhunenLoeve Expansion, etc.

The overall flow of the system can be devised as follows:

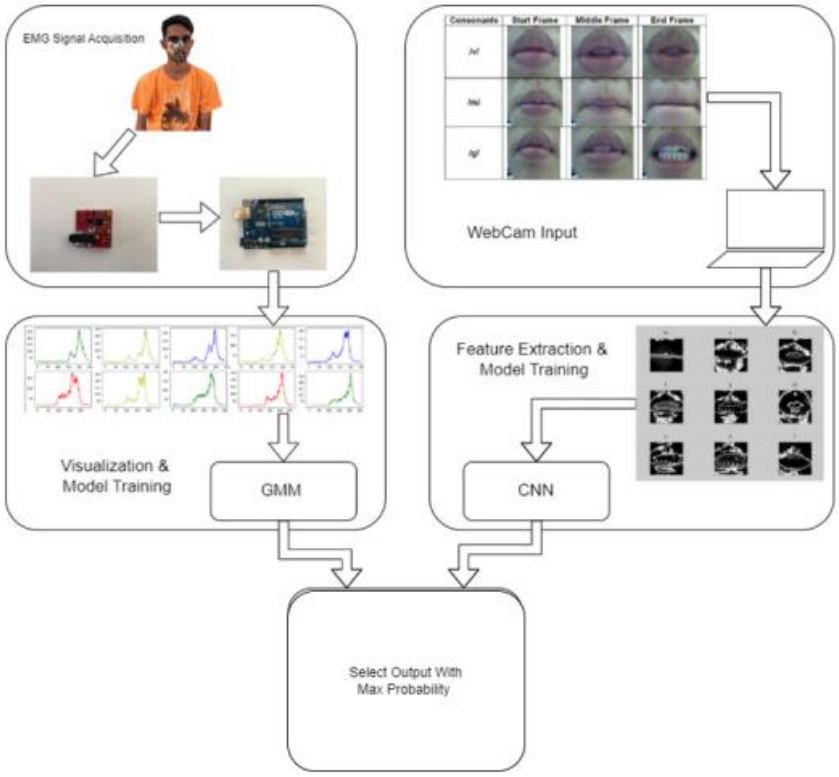

Fig. 9. Flow of the system

V. RESULTS

System obtains sequence of words spoken by the speaker from interfacing the Visual & SEMG Input. The visual based system will record the lip-movement of the user with the help of a camera and output a list of probable words the user is intending to speak. Through the list of probable words, we can select the utterance which matches our EMG system's output. So in a way, it is a dual-interrelated system. For EMG model training, a set of words were selected and its multiple utterances were recorded. For validation of the model, The accuracy for various words while using the SEMG system alone was as follows:

Word	Accuracy (Percentage %)
Heart	85
Water	55
Better	80
Continue	90

Fig. 10. Accuracy for EMG-system

These EMG signals can be further differentiated using Support Vector Machines and GMM as they have very distinct trends and/or amplitude values [2], [5], [14]. Although the GMM alone gives a good accuracy for certain words, it fails to identify correctly words which sound similar as seen above in the case of 'Water' and 'Better'. The lip-reading system has a better WER, with accuracy going up to 90% with training data, while achieving an accuracy of 53% for unseen or validation data. The accuracy and lossimprovement of the model with each epoch can be seen from the following graph as follows:

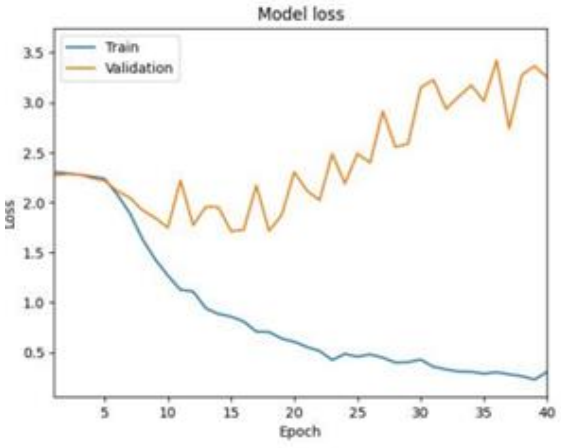

Fig. 11. Loss function of model with increasing epochs

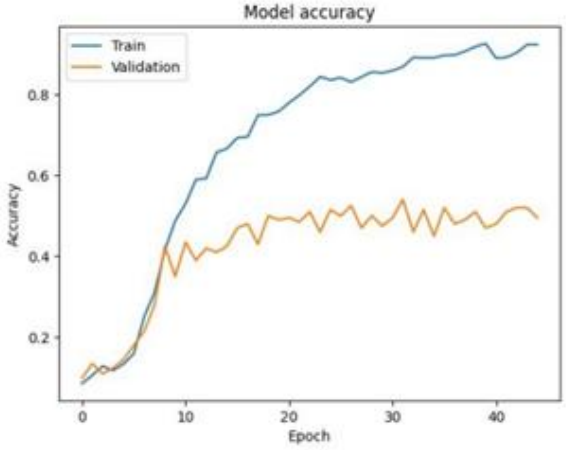

Fig. 12. Accuracy of the model with increasing epochs

VI. CONCLUSION

While the preliminary outcomes of the machine learning model may seem suboptimal, it is expected that the inclusion of additional data and more comprehensive training will enhance its accuracy. Thus, the current findings should be considered tentative and further research should aim to refine the model's learning algorithm. The most critical stage for this system is to procure EMG signals with as much high signal-to-noise ratio so it doesn't hinder the processing and analysing stage. With high accuracy recordings of neuromuscular activity, it is possible to predict silent or whispered speech. Further, in order to improve the accuracy, a visual system will be used which analyses the lip movement of the subject to predict which word the subject is uttering. Using both the systems in conjunction may also solve the problem with detecting homophones. Homophones are words which produce the same sound when uttered and thus may have similar lipmovement for some words. The muscle groups utilized in such words may be different or the lip movement may differ slightly which lets us classify them correctly to some extent. In the EMG system implemented, we are able to acquire the signal for different letters, words and sentences. Using these signals, it is possible to differentiate and identify these utterances using machine learning models and feed-back based neural networks. Another method to improve the accuracy could be using multiple electrodes to monitor different muscle groups at the same time. This would mean more data for a machine learning model to be trained on and thus more reliable results.

REFERENCES

[1] Umesh Agnihotri, AjatShatru Arora, and Atik Garg. Vowel recognition using facial movement (semg) for speech control based hci. INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH &TECHNOLOGY (IJERT) ACMEE –, 4, 2016.

[2] Adrian DC Chan, Kevin Englehart, Bernard Hudgins, and Dennis F Lovely. Myo-electric signals to augment speech recognition. Medical and Biological Engineering and Computing, 39:500–504, 2001. [3] SouheilFenghour, Daqing Chen, Kun Guo, and Perry Xiao. Lip reading sentences using deep learning with only visual cues. IEEE Access, 8:15516–215530, 2020.

[4] Yash Gondaliya, Vishaka Srinivasan, Neha Malvia, Manav Harbada, and Sheetal Jagtap. Voiceless speech recognition system. In Proceedings of the 4th International Conference on Advances in Science and Technology (ICAST2021, 2021.

[5] Mihaela Gordan, Constantine Kotropoulos, and Ioannis Pitas. Application of support vector machines classifiers to visual speech recognition. In Proceedings. International Conference on Image Processing, pages III–III, 2002.

[6] M. Janke and L. Diener. Emg-to-speech: Direct generation of speech from facial electromyographic signals. in IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(12):2375–2385, December 2017.

[7] Matthias Janke, Michael Wand, Keigo Nakamura, and Tanja Schultz. Further investigations on emg-to-speech conversion. In 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 365–368. IEEE, 2012.

[8] Sanjay Kumar, Dinesh K Kumar, Melaku Alemu, and Mark Burry. Emg-based voice recognition. In Proceedings of the 2004 Intelligent Sensors, Sensor Networks, and Information Processing Conference, 2004., pages 593–597, 2004.

[9] Alan Wee-Chung Liew and Shilin Wang. Visual Speech Recognition: Lip Segmentation and Mapping. IGI Global, 2009.

[10] L. Lu, X. Zhang, X. Xu, and Z. Wu. Homeomorphic manifold analysis: Learning motion features of image sequence for lipreading. In 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS), pages 529–532, 2015.

[11] D. Margam, R. Aralikatti, T. Sharma, and A. Thanda. arXiv preprint arXiv:1906.12170.

[12] G. S. Meltzner, J. T. Heaton, Y. Deng, G. De Luca, S. H. Roy, and J. C. Kline. Silent speech recognition as an alternative communication device for persons with laryngectomy. in IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(12):2386–2398, December 2017.

[13] Michael S Morse and Edward M O'Brien. Research summary of a scheme to ascertain the availability of speech information in the myoelectric signals of neck and head muscle using surface electrodes. Computers in biology and medicine, 16(6):399–410, 1986.

[14] MS Morse, YN Gopalan, and M Wright. Speech recognition using myoelectric signals with neural networks. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society Volume 13: 1991, pages 1877–1878, 1991.

[15] Ahmed Rekik, Achraf Ben-Hamadou, and Walid Mahdi. An adaptive approach for lip-reading using image and depth data. Multimedia Tools and Applications, 75(14):8609–8636, 2016.

[16] R Shashidhar, S Patilkulkarni, and SB Puneeth. Audio-visual speech recognition using feed-forward neural network architecture. In 2020 IEEE International Conference for Innovation in Technology (INOCON), pages 1–5, 2020.

[17] Tomoki Toda, Alan W Black, and Keiichi Tokuda. Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis. In Fifth ISCA Workshop on Speech Synthesis, 2004.

[18] Tomoki Toda, Alan W Black, and Keiichi Tokuda. Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. Speech communication, 50(3):215–227, 2008.

[19] W. C. Yau, S. P. Arjunan, and D. K. Kumar. Classification of voiceless speech using facial muscle activity and vision-based techniques. In TENCON 2008-2008 IEEE Region 10 Conference, pages 1–6, 2008.

[20] Wai Chee Yau, Dinesh Kant Kumar, and Sridhar PoosapadiArjunan. Voiceless speech recognition using dynamic visual speech features. In Proceedings of the HCSNet workshop on Use of vision in humancomputer interaction-Volume 56, pages 93–101, 2006