7. REFERENCES

1 Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. Language learning and development, 1(2), 197-234.

2 Werner, D. (1987). Disabled village children. The Hesperian Fdn, Palo Alto.

3 Long, J. S. (1908). THE SIGN LANGUAGE. A MANUAL OF SIGNS.—I. American Annals of the Deaf, 230-249.

4 Stokoe Jr, W. C. (2005). Sign language structure: An outline of the visual communication systems of the American deaf. Journal of deaf studies and deaf education, 10(1), 3-37.

5 Oghbaie, M., Sabaghi, A., Hashemifard, K., & Akbari, M. (2021). Advances and challenges in deep lip reading. arXiv preprint arXiv:2110.07879.

6 Robinson, K., & Summerfield, Q. A. (1996). Adult auditory learning and training. Ear and Hearing, 17(3), 51S-65S.

7 Houston, D. M., & Bergeson, T. R. (2014). Hearing versus listening: Attention to speech and its role in language acquisition in deaf infants with cochlear implants. Lingua, 139, 10-25.

8 Muljono, M., Saraswati, G., Winarsih, N., Rokhman, N., Supriyanto, C., & Pujiono, P. (2019). Developing BacaBicara: An Indonesian Lipreading System as an Independent Communication Learning for the Deaf and Hard-of-Hearing. International Journal of Emerging Technologies in Learning (iJET), 14(4), 44-57.

9 Isaković, L., Kovačević, T., & Dimić, N. (2016). Lip reading with deaf and hard of hearing preschool children. Thematic Collection of International Importance-Early Intervention in Special Education and Rehabilitation “, Beograd, Srbija, 2016., 195-207.

10 Berry, G. (1922). Is Adult lip-reading worth while? A Detailed study of 108 cases. The Laryngoscope, 32(9), 645-662.

11 Dodd, B. E., & Campbell, R. E. (1987). Hearing by eye: The psychology of lip-reading. Lawrence Erlbaum Associates, Inc.

12 Collen, M. F. (1994). The origins of informatics. Journal of the American Medical Informatics Association, 1(2), 91-107.

13 Sadiku, M. N., & Obiozor, C. N. (1996, November). Evolution of computer systems. In Technology-Based Re-Engineering Engineering Education Proceedings of Frontiers in Education FIE'96 26th Annual Conference (Vol. 3, pp. 1472-1474. IEEE.

14 Rojas, R., & Hashagen, U. (Eds.). (2002). The first computers: History and architectures. MIT press.

15 Campbell-Kelly, M., Aspray, W. F., Yost, J. R., Tinn, H., & Díaz, G. C. (2023). Computer: A history of the information machine. Routledge.

16 Davis, K. H., Biddulph, R., & Balashek, S. (1952). Automatic recognition of spoken digits. The Journal of the Acoustical Society of America, 24(6), 637-642.

17 Dudley, H., & Balashek, S. (1958). Automatic recognition of phonetic patterns in speech. The Journal of the Acoustical Society of America, 30(8), 721-732.

18 Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. The journal of the acoustical society of america, 26(2), 212-215.

19 Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition. IEEE Transactions on acoustics, speech, and signal processing, 23(1), 67-72.

20 Ali, M. (1976, April). Computers applied for the recognition of Hindi syllables. In ICASSP'76. IEEE International Conference on Acoustics, Speech, and Signal Processing (Vol. 1, pp. 218-221). IEEE.

21 Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. IEEE transactions on acoustics, speech, and signal processing, 26(1), 43-49.

22 Datta, A. K., & Ganguli, N. R. (1980). Automatic speech recognition in intelligence communication. IETE Journal of Research, 26(1), 82-84.

23 Paliwal, K. K., Sinha, S. S., & Agarwal, A. (1983). An Isolated Word Recognition System for Hindi Digits Using Linear Time Normalization. IETE Journal of Research, 29(1), 18-22.

24 Juang, B. H., & Rabiner, L. R. (2005). Automatic speech recognition–a brief history of the technology development. Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara, 1, 67.

25 Petajan, E. D. (1984). Automatic lipreading to enhance speech recognition (speech reading). University of Illinois at Urbana-Champaign.

26 Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

27 King, D. E. (2009). Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research, 10, 1755-1758.

28 Pearson, D. (1981). Visual communication systems for the deaf. IEEE Transactions on Communications, 29(12), 1986-1992.

29 Wark, T., Sridharan, S., & Chandran, V. (1998, August). An approach to statistical lip modelling for speaker identification via chromatic feature extraction. In Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170) (Vol. 1, pp. 123-125). IEEE.

30 Lewis, T. W., & Powers, D. M. (2000, December). Lip feature extraction using red exclusion. In Selected papers from the Pan-Sydney workshop on Visualisation-Volume 2 (pp. 61-67).

31 Skodras, E., & Fakotakis, N. (2011, May). An unconstrained method for lip detection in color images. In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1013-1016). IEEE.

32 Ghaleh, V. E. C., & Behrad, A. (2010, September). Lip contour extraction using RGB color space and fuzzy c-means clustering. In 2010 IEEE 9th International Conference on Cyberntic Intelligent Systems (pp. 1-4). IEEE.

33 Gritzman, A. D., Rubin, D. M., & Pantanowitz, A. (2015). Comparison of colour transforms used in lip segmentation algorithms. Signal, Image and Video Processing, 9(4), 947-957.

34 Viola, P., & Jones, M. (2001, December). Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001 (Vol. 1, pp. I-I). IEEE.

35 Mita, T., Kaneko, T., & Hori, O. (2005, October). Joint haar-like features for face detection. In Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 (Vol. 2, pp. 1619-1626). IEEE.

36 Wang, L., Wang, X., & Xu, J. (2010, August). Lip detection and tracking using variance based haar-like features and kalman filter. In 2010 Fifth International

Conference on Frontier of Computer Science and Technology (pp. 608-612). IEEE.

37 Puviarasan, N., & Palanivel, S. (2011). Lip reading of hearing impaired persons using HMM. Expert Systems with Applications, 38(4), 4477-4481.

38 Luettin, J., & Thacker, N. A. (1997). Speechreading using probabilistic models. Computer vision and image understanding, 65(2), 163-178.

39 Nguyen, Q. D., & Milgram, M. (2008, August). Multi features active shape models for lip contours detection. In 2008 International Conference on Wavelet Analysis and Pattern Recognition (Vol. 1, pp. 172-176). IEEE.

40 Cootes, T. F., Edwards, G. J., & Taylor, C. J. (1998). Active appearance models. In Computer Vision—ECCV'98: 5th European Conference on Computer Vision Freiburg, Germany, June 2–6, 1998 Proceedings, Volume II 5 (pp. 484-498). Springer Berlin Heidelberg.

41 Rothkrantz, L. (2017, May). Lip-reading by surveillance cameras. In 2017 Smart City Symposium Prague (SCSP) (pp. 1-6). IEEE.

42 Kass, M., Witkin, A., & Terzopoulos, D. (1988). Snakes: Active contour models. International journal of computer vision, 1(4), 321-331.

43 Kazemi, V., & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1867-1874).

44 Lee, C. G., Lee, E. S., Jung, S. T., & Lee, S. S. (2004). Design and implementation of a real-time lipreading system using PCA and HMM. Journal of Korea Multimedia Society, 7(11), 1597-1609.

45 Yao, J., & Kaifeng, Z. (2016, April). Evaluation model of the artist based on fuzzy membership to improve the principal component analysis of robust kernel. In 2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS) (pp. 322-326). IEEE.

46 Sterpu, G., & Harte, N. (2017). Towards Lipreading Sentences with Active Appearance Models. In AVSP (pp. 70-75).

47 Matthews, I., Potamianos, G., Neti, C., & Luettin, J. (2001, August). A comparison of model and transform-based visual features for audio-visual

LVCSR. In IEEE International Conference on Multimedia and Expo, 2001. ICME 2001. (pp. 210-210). IEEE Computer Society.

48 Morade, S. S., & Patnaik, S. (2014, February). Lip reading using DWT and LSDA. In 2014 IEEE International Advance Computing Conference (IACC) (pp. 1013-1018). IEEE.

49 He, J., Zhang, H., & Liu, J. Z. (2009). LDA based feature extraction method in DCT domain in lipreading. Computer Engineering and Applications, 45(32), 150-155.

50 Liang, Y., Yao, W., & Du, M. (2010, October). Feature extraction based on lsda for lipreading. In 2010 International Conference on Multimedia Technology (pp. 1-4). IEEE.

51 Almajai, I., Cox, S., Harvey, R., & Lan, Y. (2016, March). Improved speaker independent lip reading using speaker adaptive training and deep neural networks. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2722-2726). IEEE.

52 Potamianos, G., Luettin, J., & Neti, C. (2001, May). Hierarchical discriminant features for audio-visual LVCSR. In 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221) (Vol. 1, pp. 165-168). IEEE.

53 Zhou, Z., Zhao, G., & Pietikäinen, M. (2011, June). Towards a practical lipreading system. In CVPR 2011 (pp. 137-144). IEEE.

54 Zhao, G., & Pietikainen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE transactions on pattern analysis and machine intelligence, 29(6), 915-928.

55 Goldschen, A. J., Garcia, O. N., & Petajan, E. (1994, November). Continuous optical automatic speech recognition by lipreading. In Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers (Vol. 1, pp. 572-577). IEEE.

56 Shaikh, A. A., Kumar, D. K., Yau, W. C., Azemin, M. C., & Gubbi, J. (2010, October). Lip reading using optical flow and support vector machines. In 2010 3Rd international congress on image and signal processing (Vol. 1, pp. 327-330). IEEE.

57 Cappelletta, L., & Harte, N. (2011, August). Viseme definitions comparison for visual-only speech recognition. In 2011 19th European Signal Processing Conference (pp. 2109-2113). IEEE.

58 Ma, X., Yan, L., & Zhong, Q. (2016, July). Lip feature extraction based on improved jumping-snake model. In 2016 35th Chinese Control Conference (CCC) (pp. 6928-6933). IEEE.

59 Hao, M., Mamut, M., Yadikar, N., Aysa, A., & Ubul, K. (2020). A Survey of Research on Lipreading Technology. IEEE Access.

60 Watanabe, T., Katsurada, K., & Kanazawa, Y. (2016, November). Lip reading from multi view facial images using 3D-AAM. In Asian Conference on Computer Vision (pp. 303-316). Springer, Cham.

61 Petajan, E., Bischoff, B., Bodoff, D., & Brooke, N. M. (1988, May). An improved automatic lipreading system to enhance speech recognition. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 19-25).

62 Brahme, A., & Bhadade, U. (2017, November). Marathi digit recognition using lip geometric shape features and dynamic time warping. In TENCON 2017-2017 IEEE Region 10 Conference (pp. 974-979). IEEE.

63 Bennani, Y., & Gallinari, P. (1991, April). On the use of TDNN-extracted features information in talker identification. In Acoustics, Speech, and Signal Processing, IEEE International Conference on (pp. 385-388). IEEE Computer Society.

64 Bregler, C., & Konig, Y. (1994, April). " Eigenlips" for robust speech recognition. In Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing (Vol. 2, pp. II-669). IEEE.

65 Sujatha, P., & Krishnan, M. R. (2012, February). Lip feature extraction for visual speech recognition using Hidden Markov Model. In 2012 International Conference on Computing, Communication and Applications (pp. 1-5). IEEE.

66 Thangthai, K., Bear, H. L., & Harvey, R. (2018). Comparing phonemes and visemes with DNN-based lipreading. arXiv preprint arXiv:1805.02924.

67 Zhang, K., Zhang, Z., Wang, H., Li, Z., Qiao, Y., & Liu, W. (2017). Detecting faces using inside cascaded contextual CNN. In Proceedings of the IEEE international conference on computer vision (pp. 3171-3179).

68 Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... & Darrell, T. (2014). Caffe: Convolutional Architecture for Fast Feature Embedding. arXiv preprint arXiv:1408.5093

69 Wand, M., Koutník, J., & Schmidhuber, J. (2016, March). Lipreading with long short-term memory. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6115-6119). IEEE.

70 Wand, M., & Schmidhuber, J. (2017). Improving speaker-independent lipreading with domain-adversarial training. arXiv preprint arXiv:1708.01565.

71 Wand, M., Schmidhuber, J., & Vu, N. T. (2018, April). Investigations on end-to-end audiovisual fusion. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3041-3045). IEEE.

72 Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527-1554. https://doi.org/10.1162/neco.2006.18.7.1527

73 Sui, C., Bennamoun, M., & Togneri, R. (2015). Listening with your eyes: Towards a practical visual speech recognition system using deep boltzmann machines. In Proceedings of the IEEE International Conference on Computer Vision (pp. 154-162).

74 Petridis, S., Li, Z., & Pantic, M. (2017, March). End-to-end visual speech recognition with LSTMs. In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 2592-2596). IEEE.

75 LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278-2324.

76 Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

77 Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

78 Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

79 He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

80 Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint arXiv:1704.04861*

81 Ji, S., Xu, W., Yang, M., & Yu, K. (2012). 3D convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 35(1), 221-231.

82 Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

83 Graves, A., Mohamed, A.-R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6645-6649.

84 Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11), 2673-2681

85 Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). A Critical Review of Recurrent Neural Networks for Sequence Learning. arXiv preprint arXiv:1506.00019.

86 Nishida, S. (1986). Speech recognition enhancement by lip information. ACM SIGCHI Bulletin, 17(4), 198-204.

87 Sejnowski, T. J., & JOHNS HOPKINS UNIV BALTIMORE MD. (1988). Massively-Parallel Architectures for Automatic Recognition of Visual Speech Signals.

88 Sejnowski, T. J., Yuhas, B., Goldstein, M., & Jenkins, R. (1989). Combining visual and acoustic speech signals with a neural network improves intelligibility. Advances in neural information processing systems, 2.

89 Goldschen, A. J., Garcia, O. N., & Petajan, E. D. (1997). Continuous automatic speech recognition by lipreading. In Motion-Based recognition (pp. 321-343). Dordrecht: Springer Netherlands.

90 Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., ... & Mashari, A. (2000). Audio visual speech recognition.

91 Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., & Ogata, T. (2014, September). Lipreading using convolutional neural network. In Interspeech (Vol. 1, p. 3).

92 Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., & Ogata, T. (2015). Audio-visual speech recognition using deep learning. Applied intelligence, 42, 722-737.

93 Garg, A., Noyola, J., & Bagadia, S. (2016). Lip reading using CNN and LSTM. Technical report, Stanford University, CS231 n project report.

94 Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition.

95 Li, Y., Takashima, Y., Takiguchi, T., & Ariki, Y. (2016, June). Lip reading using a dynamic feature of lip images and convolutional neural networks. In 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS) (pp. 1-6). IEEE.

96 Chung, J. S., & Zisserman, A. (2017). Out of time: automated lip sync in the wild. In Asian conference on computer vision (pp. 251-263). Springer, Cham.

97 Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90.

98 Saitoh, T., Zhou, Z., Zhao, G., & Pietikäinen, M. (2017). Concatenated frame image based CNN for visual speech recognition. In Asian Conference on Computer Vision (pp. 277-289). Springer, Cham.

99 Lin, M., Chen, Q., & Yan, S. (2013). Network in network. arXiv preprint arXiv:1312.4400.

100 Zhang, X., Gong, H., Dai, X., Yang, F., Liu, N., & Liu, M. (2019, July). Understanding pictograph with facial features: end-to-end sentence-level lip reading of Chinese. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 9211-9218).

101 Ji, S., Xu, W., Yang, M., & Yu, K. (2012). 3D convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 35(1), 221-231.

102 Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision (pp. 4489-4497).

103 Tran, D., Ray, J., Shou, Z., Chang, S. F., & Paluri, M. (2017). Convnet architecture search for spatiotemporal feature learning. arXiv preprint arXiv:1708.05038.

104 Qiu, Z., Yao, T., & Mei, T. (2017). Learning spatio-temporal representation with pseudo-3d residual networks. In proceedings of the IEEE International Conference on Computer Vision (pp. 5533-5541).

105 Assael, Y. M., Shillingford, B., Whiteson, S., & De Freitas, N. (2016). Lipnet: End-to-end sentence-level lipreading. arXiv preprint arXiv:1611.01599.

106 Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural networks, 18(5-6), 602-610.

107 Fung, I., & Mak, B. (2018, April). End-to-end low-resource lip-reading with maxout CNN and LSTM. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2511-2515). IEEE.

108 Torfi, A., Iranmanesh, S. M., Nasrabadi, N., & Dawson, J. (2017). 3d convolutional neural networks for cross audio-visual matching recognition. IEEE Access, 5, 22081-22091.

109 Chung, J. S., & Zisserman, A. (2017). Lip reading in the wild. In Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13 (pp. 87-103). Springer International Publishing.

110 Xu, K., Li, D., Cassimatis, N., & Wang, X. (2018, May). LCANet: End-to-end lipreading with cascaded attention-CTC. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) (pp. 548-555). IEEE.

111 Srivastava, R. K., Greff, K., & Schmidhuber, J. (2015). Training very deep networks. Advances in neural information processing systems, 28.

112 Yang, S., Zhang, Y., Feng, D., Yang, M., Wang, C., Xiao, J., ... & Chen, X. (2019, May). LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019) (pp. 1-8). IEEE.

113 Stafylakis, T., & Tzimiropoulos, G. (2018, April). Deep word embeddings for visual speech recognition. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4974-4978). IEEE.

114 Margam, D. K., Aralikatti, R., Sharma, T., Thanda, A., Roy, S., & Venkatesan, S. M. (2019). LipReading with 3D-2D-CNN BLSTM-HMM and word-CTC models. arXiv preprint arXiv:1906.12170.

115 Xiao, J., Yang, S., Zhang, Y., Shan, S., & Chen, X. (2020, November). Deformation flow based two-stream network for lip reading. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (pp. 364-370). IEEE.

116 Luo, M., Yang, S., Shan, S., & Chen, X. (2020, November). Pseudo-convolutional policy gradient for sequence-to-sequence lip-reading. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (pp. 273-280). IEEE.

117 Zhang, Y., Yang, S., Xiao, J., Shan, S., & Chen, X. (2020, November). Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (pp. 356-363). IEEE.

118 Zhao, X., Yang, S., Shan, S., & Chen, X. (2020, November). Mutual information maximization for effective lip reading. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (pp. 420-427). IEEE.

119 Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011, January). Multimodal deep learning. In ICML.

120 Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., & Pantic, M. (2018, April). End-to-end audiovisual speech recognition. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 6548-6552). IEEE.

121 Petridis, S., Wang, Y., Li, Z., & Pantic, M. (2017). End-to-end audiovisual fusion with LSTMs. arXiv preprint arXiv:1709.04343.

122 Petridis, S., Wang, Y., Li, Z., & Pantic, M. (2017). End-to-end multi-view lipreading. arXiv preprint arXiv:1709.00443.

123 Saitoh, T., Morishita, K., & Konishi, R. (2008, December). Analysis of efficient lip reading method for various languages. In 2008 19th International Conference on Pattern Recognition (pp. 1-4). IEEE.

124 Movellan, J. (1994). Visual speech recognition with stochastic networks. Advances in neural information processing systems, 7.

125 Vanegas, O., Tokuda, K., & Kitamura, T. (1999, October). Location normalization of HMM-based lip-reading: experiments for the M2VTS database. In Proceedings 1999 International Conference on Image Processing (Cat. 99CH36348) (Vol. 2, pp. 343-347). IEEE.

126 Messer, K., Matas, J., Kittler, J., Luettin, J., & Maitre, G. (1999, March). XM2VTSDB: The extended M2VTS database. In Second international conference on audio and video-based biometric person authentication (Vol. 964, pp. 965-966).

127 Matthews, I., Cootes, T. F., Bangham, J. A., Cox, S., & Harvey, R. (2002). Extraction of visual features for lipreading. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(2), 198-213.

128 Patterson, E. K., Gurbuz, S., Tufekci, Z., & Gowdy, J. N. (2002, May). CUAVE: A new audio-visual database for multimodal human-computer interface research. In 2002 IEEE International conference on acoustics, speech, and signal processing (Vol. 2, pp. II-2017). IEEE.

129 Bailly-Bailliére, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariéthoz, J., ... & Thiran, J. P. (2003). The BANCA database and evaluation protocol. In Audio-and Video-Based Biometric Person Authentication: 4th International Conference, AVBPA 2003 Guildford, UK, June 9–11, 2003 Proceedings 4 (pp. 625-638). Springer Berlin Heidelberg.

130 Ortega, A., Sukno, F., Lleida, E., Frangi, A. F., Miguel, A., Buera, L., & Zacur, E. (2004, May). AV@ CAR: A Spanish Multichannel Multimodal Corpus for In-Vehicle Automatic Audio-Visual Speech Recognition. In LREC.

131 Lee, B., Hasegawa-Johnson, M., Goudeseune, C., Kamdar, S., Borys, S., Liu, M., & Huang, T. (2004). AVICAR: Audio-visual speech corpus in a car environment. In Eighth International Conference on Spoken Language Processing.

132 Huang, J., Potamianos, G., Connell, J., & Neti, C. (2004). Audio-visual speech recognition using an infrared headset. Speech Communication, 44(1-4), 83-96.

133 Fox, N. A., O'Mullane, B. A., & Reilly, R. B. (2005, July). VALID: A new practical audio-visual database, and comparative results. In International conference on audio-and video-based biometric person authentication (pp. 777-786). Berlin, Heidelberg: Springer Berlin Heidelberg.

134 Cox, S. J., Harvey, R. W., Lan, Y., Newman, J. L., & Theobald, B. J. (2008, September). The challenge of multispeaker lip-reading. In AVSP (pp. 179-184).

135 Lucey, P., Potamianos, G., & Sridharan, S. (2008). Patch-based analysis of visual speech from multiple views. In Proceedings of the International Conference on Auditory-Visual Speech Processing 2008 (pp. 69-74). AVISA.

136 Tamura, S., Miyajima, C., Kitaoka, N., Yamada, T., Tsuge, S., Takiguchi, T., ... & Nakamura, S. (2010). CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition. In Auditory-Visual Speech Processing 2010.

137 Pass, A., Zhang, J., & Stewart, D. (2010, September). An investigation into features for multi-view lipreading. In 2010 IEEE International Conference on Image Processing (pp. 2417-2420). IEEE.

138 Chitu, A. G., Driel, K., & Rothkrantz, L. J. (2010). Automatic lip reading in the Dutch language using active appearance models on high speed recordings. In Text, Speech and Dialogue: 13th International Conference, TSD 2010, Brno, Czech Republic, September 6-10, 2010. Proceedings 13 (pp. 259-266). Springer Berlin Heidelberg.

139 Estellers, V., & Thiran, J. P. (2011, August). Multipose audio-visual speech recognition. In 2011 19th European Signal Processing Conference (pp. 1065-1069). IEEE.

140 Igras, M., Ziółko, B., & Jadczyk, T. (2012). Audiovisual database of Polish speech recordings. Studia Informatica, 33(2B), 163-172.

141 Estival, D., Cassidy, S., Cox, F., & Burnham, D. (2014). AusTalk: an audio-visual corpus of Australian English.

142 Zhao, G., Barnard, M., & Pietikainen, M. (2009). Lipreading with local spatiotemporal descriptors. IEEE Transactions on Multimedia, 11(7), 1254-1265.

143 Petridis, S., Shen, J., Cetin, D., & Pantic, M. (2018, April). Visual-only recognition of normal, whispered and silent speech. In 2018 ieee international conference on acoustics, speech and signal processing (icassp) (pp. 6219-6223). IEEE.

144 Pei, Y., Kim, T. K., & Zha, H. (2013). Unsupervised random forest manifold alignment for lipreading. In Proceedings of the IEEE International Conference on Computer Vision (pp. 129-136).

145 Petridis, S., & Pantic, M. (2016, March). Deep complementary bottleneck features for visual speech recognition. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2304-2308). IEEE.

146 Hu, D., & Li, X. (2016). Temporal multimodal learning in audiovisual speech recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3574-3582).

147 Zue, V., Seneff, S., & Glass, J. (1990). Speech database development at MIT: TIMIT and beyond. Speech communication, 9(4), 351-356.

148 Fu, Y., Zhou, X., Liu, M., Hasegawa-Johnson, M., & Huang, T. S. (2007, September). Lipreading by locality discriminant graph. In 2007 IEEE International Conference on Image Processing (Vol. 3, pp. III-325). IEEE.

149 Papandreou, G., Katsamanis, A., Pitsikalis, V., & Maragos, P. (2009). Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 17(3), 423-435.

150 Rahmani, M. H., & Almasganj, F. (2017, April). Lip-reading via a DNN-HMM hybrid system using combination of the image-based and model-based features. In 2017 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA) (pp. 195-199). IEEE.

151 Ninomiya, H., Kitaoka, N., Tamura, S., Iribe, Y., & Takeda, K. (2015, September). Integration of deep bottleneck features for audio-visual speech recognition. In Interspeech (pp. 563-567).

152 Rothkrantz, L. J. M. (2012). Automatic Visual Speech Recognition. Speech Enhancement, Modeling and Recognition-Algorithms and Applications, ISBN: 978-953-51-0291-5.

153 Sui, C., Togneri, R., & Bennamoun, M. (2015, April). Extracting deep bottleneck features for visual speech recognition. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1518-1522). IEEE.

154 Sui, C., Togneri, R., & Bennamoun, M. (2017). A cascade gray-stereo visual feature extraction method for visual and audio-visual speech recognition. Speech Communication, 90, 26-38.

155 Sanderson, C. (2002). The vidtimit database.

156 Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America, 120(5), 2421-2424.

157 Anina, I., Zhou, Z., Zhao, G., & Pietikäinen, M. (2015, May). Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis. In 2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG) (Vol. 1, pp. 1-5). IEEE.

158 Rekik, A., Ben-Hamadou, A., & Mahdi, W. (2014). A new visual speech recognition approach for RGB-D cameras. In *Image Analysis and Recognition: 11th International Conference, ICIAR 2014, Vilamoura, Portugal, October 22-24, 2014, Proceedings, Part II 11* (pp. 21-28). Springer International Publishing.

159 Chung, J. S., & Zisserman, A. (2017). Lip reading in the wild. In Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13 (pp. 87-103). Springer International Publishing.

160 Yang, S., Zhang, Y., Feng, D., Yang, M., Wang, C., Xiao, J., ... & Chen, X. (2019, May). LRW-1000: A naturally distributed large-scale benchmark for lip reading in the wild. In 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019) (pp. 1-8). IEEE.

161 Son Chung, J., Senior, A., Vinyals, O., & Zisserman, A. (2017). Lip reading sentences in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6447-6456).

162 Afouras, T., Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2018). Deep audio-visual speech recognition. IEEE transactions on pattern analysis and machine intelligence, 44(12), 8717-8727.

163 Afouras, T., Chung, J. S., & Zisserman, A. (2018). LRS3-TED: a large-scale dataset for visual speech recognition. arXiv preprint arXiv:1809.00496.

164 Son, J. S., & Zisserman, A. (2017, September). Lip reading in profile. In Proc. Brit. Mach. Vis. Conf.(BMVC) (pp. 1-11). Sep.

165 Stafylakis, T., & Tzimiropoulos, G. (2017). Combining residual networks with LSTMs for lipreading. arXiv preprint arXiv:1703.04105.

166 Stafylakis, T., & Tzimiropoulos, G. (2018, April). Deep word embeddings for visual speech recognition. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4974-4978). IEEE.

167 Wu, P., Liu, H., Li, X., Fan, T., & Zhang, X. (2016). A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion. IEEE Transactions on Multimedia, 18(3), 326-338.

168 Lee, D., Lee, J., & Kim, K. E. (2017). Multi-view automatic lip-reading using neural network. In Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13 (pp. 290-302). Springer International Publishing.

169 Tailor, J. H., & Shah, D. B. (2015). Review on Speech Recognition System for Indian Languages. International Journal of Computer Applications, 119(2).

170 Ali, M. (1976, April). Computers applied for the recognition of Hindi syllables. In ICASSP'76. IEEE International Conference on Acoustics, Speech, and Signal Processing (Vol. 1, pp. 218-221). IEEE.

171 Paliwal, K. K., Sinha, S. S., & Agarwal, A. (1983). An Isolated Word Recognition System for Hindi Digits Using Linear Time Normalization. IETE Journal of Research, 29(1), 18-22.

172 Patil, A., More, P., & Sasikumar, M. (2019). Incorporating finer acoustic phonetic features in lexicon for Hindi language speech recognition. Journal of Information and Optimization Sciences, 40(8), 1731-1739.

173 Faruquie, T. A., Neti, C., Rajput, N., Subramaniam, L. V., & Verma, A. (2000, July). Translingual visual speech synthesis. In 2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532) (Vol. 2, pp. 1089-1092). IEEE.

174 Kandagal, A. P., & Udayashankara, V. (2017). Visual Speech Recognition Based on Lip Movement for Indian Languages. Int. J. Comput. Intell. Res, 13, 2029-2041.

175 Nandini, M. S., Nagavi, T. C., & Bhajantri, N. U. (2019, March). Deep Weighted Feature Descriptors for Lip Reading of Kannada Language. In 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN) (pp. 978-982). IEEE.

176 Patil, M. S., Chickerur, S., Meti, A., Nabapure, P. M., Mahindrakar, S., Naik, S., & Kanyal, S. (2019). LSTM Based Lip Reading Approach for Devanagiri Script.

177 Rudregowda, S., Patil Kulkarni, S., HL, G., Ravi, V., & Krichen, M. (2023, March). Visual speech recognition for kannada language using vgg16 convolutional neural network. In Acoustics (Vol. 5, No. 1, pp. 343-353). MDPI.

178 Akhter, N. (2016). A viseme recognition system using lip curvature and neural networks to detect Bangla vowels (Doctoral dissertation, BRAC Univeristy).

179 Census India 2011
http://www.censusindia.gov.in/

180 Hemakumar, G., & Punitha, P. (2013). Speech recognition technology: a survey on Indian languages. International Journal of Information Science and Intelligent System, 2(4), 1-38.

181 Parikh, R. B., & Joshi, H. (2020). Gujarati Speech Recognition–A Review. no, 549, 6.

182 Patel, J., & Nandurbarkar, A. (2015). Development and implementation of algorithm for speaker recognition for gujarati language. International Research Journal of Engineering and Technology, 2(2), 444-448

183 Vijayendra, A. D., & Thakar, V. K. (2016). Neural network based Gujarati speech recognition for dataset collected by in-ear microphone. Procedia computer science, 93, 668-675.

184 Valaki, S., & Jethva, H. (2017, March). A hybrid HMM/ANN approach for automatic Gujarati speech recognition. In 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS) (pp. 1-5). IEEE.

185 Tailor, J. H., & Shah, D. B. (2018). HMM-based lightweight speech recognition system for gujarati language. In Information and Communication Technology

for Sustainable Development: Proceedings of ICT4SD 2016, Volume 2 (pp. 451-461). Springer Singapore.

186 Raval, D., Pathak, V., Patel, M., & Bhatt, B. (2020, December). End-to-End Automatic Speech Recognition for Gujarati. In Proceedings of the 17th International Conference on Natural Language Processing (ICON) (pp. 409-419).

187 Pandit, P., Bhatt, S., & Makwana, P. (2014). Automatic speech recognition of Gujarati digits using artificial neural network. In Proceedings of 19th Annual Cum 4th International Conference of GAMS On Advances in Mathematical Modelling to Real World Problems (pp. 141-146).

188 Tailor, J. H., Rakholia, R., Saini, J. R., & Kotecha, K. (2022). Deep Learning Approach for Spoken Digit Recognition in Gujarati Language. International Journal of Advanced Computer Science and Applications, 13(4).

189 Almajai, I., Cox, S., Harvey, R., & Lan, Y. (2016, March). Improved speaker independent lip reading using speaker adaptive training and deep neural networks. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2722-2726). IEEE.

190 Fernandez-Lopez, A., & Sukno, F. M. (2017). Automatic viseme vocabulary construction to enhance continuous lip-reading. arXiv preprint arXiv:1704.08035.

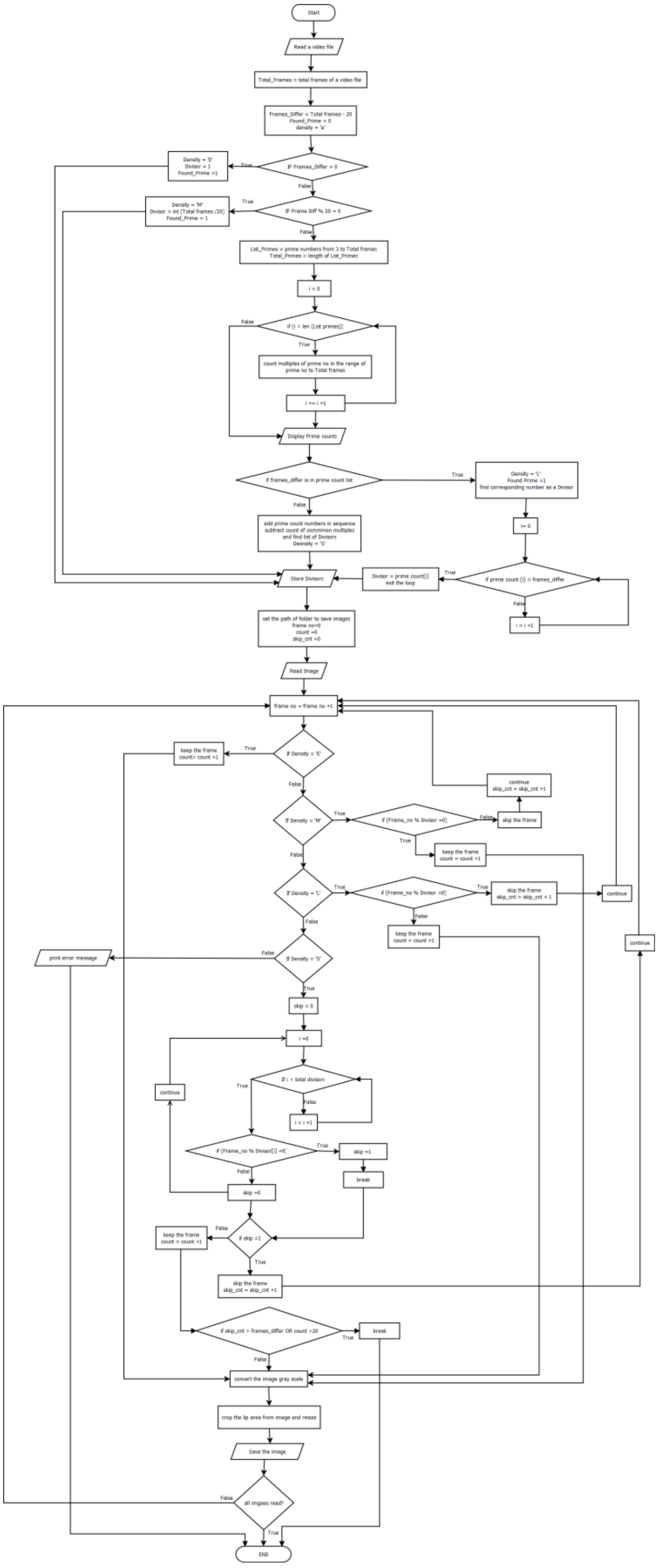
191 Bear, H. L., Harvey, R. W., Theobald, B. J., & Lan, Y. (2014). Which phoneme-to-viseme maps best improve visual-only computer lip-reading?. In Advances in Visual Computing: 10th International Symposium, ISVC 2014, Las Vegas, NV, USA, December 8-10, 2014, Proceedings, Part II 10 (pp. 230-239). Springer International Publishing.

192 Stafylakis, T., & Tzimiropoulos, G. (2017). Combining residual networks with LSTMs for lipreading. arXiv preprint arXiv:1703.04105.

193 Yang, S., Zhang, Y., Feng, D., Yang, M., Wang, C., Xiao, J., ... & Chen, X. (2019, May). LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019) (pp. 1-8). IEEE.

194 Zhou, Z., Hong, X., Zhao, G., & Pietikäinen, M. (2013). A compact representation of visual speech data using latent variables. IEEE transactions on pattern analysis and machine intelligence, 36(1), 1-1.

ANNEXURE –ViLiDEx Flowchart



Start
Read a video file
Total_Frames = total frames of a video file
Frames_Differ = Total frames - 20
Found_Prime = 0
density = 'e'
IF Frames_Differ = 0
True
Density = 'E'
Divisor = 1
Found_Prime =1
False
IF Frame Diff % 20 = 0
True
Density = 'M'
Divisor = int (Total frames /20)
Found_Prime = 1
False
List_Primes = prime numbers from 3 to Total frames
Total_Primes = length of List_Primes
i = 0
if (i < len (List primes))
False
True
count multiples of prime no in the range of prime no to Total frames
i += i +1
Display Prime counts
if frames_differ is in prime count list
True
Density = 'L'
Found Prime =1
find corresponding number as a Divisor
i= 0
if prime count [i] = frames_differ
True
Divisor = prime count[i]
exit the loop
False
i = i +1
False
add prime count numbers in sequence
subtract count of commmon multiples
and find list of Divisors
Deisnity = 'S'
Store Divisors
set the path of folder to save images
frame no=0
count =0
skip_cnt =0
Read Image
frame no = frame no +1
If Density = 'E'
True
keep the frame
count= count +1
False
If Density = 'M'
True
if (Frame_no % Divisor =0)
False
skip the frame
continue
skip_cnt = skip_cnt +1
True
keep the frame
count = count +1
False
If Density = 'L'
True
if (Frame_no % Divisor =0)
True
skip the frame
skip_cnt = skip_cnt + 1
continue
False
keep the frame
count = count +1
continue
False
If Density = 'S'
False
print error message
True
skip = 0
i =0
If i < total divisors
True
False
i = i +1
continue
if (Frame_no % Divisor[i] =0)
True
skip =1
break
False
skip =0
if skip =1
False
keep the frame
count = count +1
True
skip the frame
skip_cnt = skip_cnt +1
if skip_cnt > frames_differ OR count >20
True
break
False
convert the image gray scale
crop the lip area from image and resize
Save the image
all imgaes read?
False
True
END