INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

PRAV: A Phonetically Rich Audio Visual Corpus

Abhishek Narwekar[1], *Prasanta Kumar Ghosh*[2]

[1]Dept of Computer Science, University of Illinois at Urbana Champaign, IL, USA 61801
[2]Dept of Electrical Engineering, Indian Institute of Science, Bangalore, India 560012

an41@illinois.edu, prasantg@ee.iisc.ernet.in

Abstract

This paper describes the acquisition of PRAV, a phonetically rich audio-visual Corpus. The PRAV Corpus contains audio as well as visual recordings of 2368 sentences from the TIMIT corpus each spoken by four subjects, making it the largest audio-visual corpus in the literature in terms of the number of sentences per subject. Visual features, comprising the coordinates of points along the contour of the subjects lips, have been extracted for the entire PRAV Corpus using the Active Appearance Models (AAM) algorithm and have been made available along with the audio and video recordings. The subjects being Indian makes PRAV an ideal resource for audio-visual speech study with non-native English speakers. Moreover, this paper describes how the large number of sentences per subject makes the PRAV Corpus a significant dataset by highlighting its utility in exploring a number of potential research problems including visual speech synthesis and perception studies.

Index Terms: audio visual dataset, audio visual speech synthesis

1. Introduction

Audio-visual data provides rich multimodal information about both linguistic and paralinguistic modes in a face-to-face communication. Audio and visual information are often complementary in nature. Several audio-visual datasets have been created in the past to develop engineering models exploiting this rich information for a variety of technological solutions.

One of the earliest problems reported using the audio-visual data was phoneme classification by supplementing the audio with the visual data which is typically used to capture features from the lip region [1] while subjects speak various phonemes [2]. In addition to capturing lip movement, dynamics of internal articulators such as tongue, velum have been captured using real time magnetic resonance imaging (rtMRI) data in the midsagittal upper airway and used as information complementary to audio [3]. The task of phoneme classification was extended to audio-visual speech recognition (AVSR) using several audio-visual corpora including GRID [4] and others. For training an audio-visual speech recognizer, it is typically required to have a large audio-visual dataset rich with phonetic contexts as well as speaker variability. There are a number of such large audio-visual datasets available in the literature. For example, the IBM's ViaVoice [1] is a dataset with 290 speakers and over 50 hours of recording, containing utterances from the IBM ViaVoice Training Transcripts. Some Corpora such as GRID [4] and VidTIMIT [5] have significant speaker diversity with 34 and 43 speakers respectively. But they have little phonetic and lexical diversity, due to either few sentences or sentences that are restrictive in nature. For instance, VidTIMIT has only 10 sentences per speaker, while GRID has sentences of the form "put red at G9 now". Similar corpora have been used for AVSR in "challenging" environments such as in a moving car or through cameras in phones and laptops; for example, MOBIO [6], has sentences in a question-answer format.

The task of emotion detection has been addressed with a variety of audio-visual datasets, both English [7] and non-English [8]. The datasets used in these works differ in the nature of the stimulus provided to the subjects. For instance, some works attempt to capture emotion from interaction between subjects in scripted plays [9].

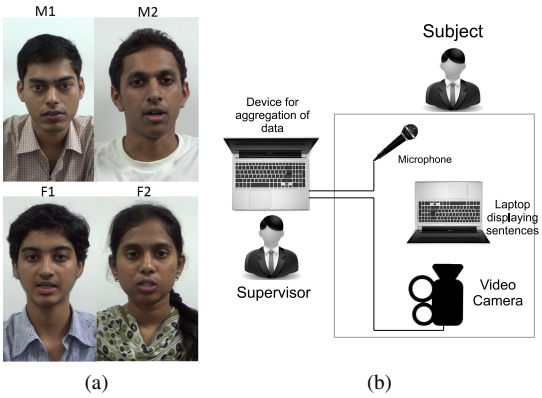
In human-machine communication, the interaction could happen through audio-visual modes. Datasets such as AVLetters [10] have recordings of individual alphabets or strings of numbers used in an interactive voice response (IVR) system.

Audio visual speech synthesis is another area of research where audio-visual data plays a crucial role. There have been various approaches to synthesize audio visual speech in the past. Visual speech synthesis has been traditionally approached as a problem of modelling the face as a 3D object [11]. The dataset in [12] has 3D data of markers on and around the subjects' faces during speech and is useful for creating a face model. The mapping of phonemes to visemes is an important sub-problem in a two-phase visual speech synthesis [13]. This is critical for another line of work on 2D audio-visual speech synthesis [14] where visemes are extracted from the video and are integrated with the appropriate phonemes.

The problem of speaker detection [15] is inherently an important one while distinguishing twins, as done using databases such as ND-Twins [16]. Datasets such as XM2VTS [17] are also well-suited for user authentication.

In this article, we present PRAV Corpus, a large phonetically rich audio-visual dataset with over 9000 sentences uttered by four subjects, totaling to about 5 hours per subject. Only a few audio-visual datasets such as the IBM ViaVoice [1] provide such large volume of data. However, ViaVoice has an average duration of around 10 minutes per subject. Similarly, the OuluVS2 corpus [18] too contains 52 speakers, but with limited utterances per speaker. The AusTalk corpus [19] also contains recordings from a large number of Australian speakers for a variety of tasks such as reciting digit strings, stories and spontaneous speech. In contrast to a large number of subjects with small per-subject data as in ViaVoice and OuluVS2, PRAV is rich with phonetic and lexical diversity on a per subject basis and is generic enough to perform most of the tasks described above. As opposed to AusTalk, PRAV contains sentences from the TIMIT corpus, making it more phonetically balanced. Since subjects in the PRAV Corpus are Indian, it is also useful for audio-visual study of speech from non-native English speakers. Along with the audio and video recordings in PRAV Corpus, we also provide supplementary data such as annotations for sentence boundaries and lip features extracted from the recordings of the entire database for further studies.

http://dx.doi.org/10.21437/Interspeech.2017-242



(a) (b)

Figure 1: *(a) Snapshots of each subject from the recordings in PRAV Corpus, (b) A schematic of the top view of the apparatus used for recording the PRAV Corpus*

2. Dataset Acquisition

PRAV Corpus consists of audio visual recordings of 2368 phonetically balanced sentences; these are the unique sentences in the TIMIT corpus [20]. Each sentence was spoken by 4 subjects (2 male - M1, M2 - and 2 female - F1, F2), all in the age range 20 and 22 years. All four subjects were undergraduate students at the time of recording. The native languages of M1, M2, F1 and F2 were Marathi, Kannada, Malayalam and Tamil respectively. The medium of instruction in their university was English, as a result of which the subjects were fluent in reading, writing and speaking English. A snapshot of each of the subjects is shown in Figure 1(a).

The total duration of all recorded videos is nearly 20 hours. To the best of our knowledge, PRAV has the largest number of sentences per subject among all audio-visual datasets in the literature. The total duration of recording, excluding the silence regions between sentences, for four subjects is 6697, 7411, 8967 and 7318 seconds for M1, M2, F1 and F2 respectively. A copy of the dataset can be obtained from http://spire.ee.iisc.ernet.in/spire/database.php

The recording was done in a well illuminated anechoic chamber. The video was recorded at a frame rate of 25 frames/second using a Sony Handycam – model HDR-CX280E with 29.8mm wide-angle lens and 8.9 megapixels still picture resolution at an aspect ratio of 16:9 (image dimensions: 1280x720). The distance between the camera and the subject was maintained constant at 3.3 feet by fixing the position of the seat for the subject, and varying only the height of the camera in order to capture the face of the subject completely. A light backdrop was used for the video recording. A top-view schematic for the recording apparatus is shown in Figure 1(b).

The audio was recorded at a sampling frequency of 16 kHz and a bit depth of 16 bits using a Behringer Single Diaphragm Condenser Microphone, model B-1, and was stored uncompressed in the `.wav` format. Praat [21] was used for recording and storing the audio from the microphone. Each subject was instructed about the functioning of the setup and allowed to practise before the actual recordings in order to familiarize them with the protocol.

The recordings were done in batches of 100 sentences, which took around 7 to 10 minutes to finish. This was done to ensure that the subjects weren't tired of speaking and to maintain the quality of the audio and video recordings. A minimum of 5 minutes was given as a rest-time between two batches of sentences, during which the subjects were allowed to rest and relax. The sentences were displayed one per slide on a laptop computer placed between the subject and below the stand of the camera in such way that it did not block the camera view (refer Figure 1(b)). The subjects were instructed to pause for approximately one second after uttering of each sentence and navigated the slides themselves. The subjects were monitored for stuttering and for deviation of the utterance from the displayed sentences. In case of such errors, the subjects were signalled at the end of each sentence to repeat the entire sentence till it was spoken correctly. In order to remove the incorrect utterances, we annotated all the correct utterances in the database by delineating their start and end points using Audacity [22].

Two different apparatus were used for recording the audio and the video, owing to the superior quality of the audio from the microphone. Synchronization of the audio and video streams was an important issue, as it was not possible to start recordings of both streams simultaneously. We used the cross correlation between the two streams to estimate the delay between them which was used for synchronization.

3. Tool for Lip Feature Extraction

3.1. Active Appearance Model (AAM)

Features that represent lip movement are often used in several applications that use audio-visual corpora. To facilitate exploratory work using the PRAV dataset, we extract lip features from the video frames. The algorithm developed for lip feature extraction is based on the Active Appearance Model (AAM) [23]. The block diagram in Figure 2 describes the various steps required to extract lip features from the video frames of the PRAV Corpus. Selection and annotation of the training images is a key step in the lip feature extraction. This is followed by the extraction of lip features from a test image involving preprocessing the image, appropriate initialization, execution of the AAM algorithm and post-processing to detect and rectify erroneous features. We use the code for IC AAM [24] as a template, and build upon it. We had to overcome certain shortcomings of the IC AAM in order to extract lip features effectively in the PRAV Corpus. For this purpose, we have described these challenges and our solutions for the same in the following subsections. Along with the dataset, we provide the extracted lip features obtained using AAM and the code to do the same.

Training Images: In order to prepare the training data for AAM, we manually select 19 video frames from the recordings of each subject corresponding to various lip shapes in different phonetic contexts and convert them to gray scale images. Ideally, the training should be performed on visemes. Nevertheless, since our choices of training images are phoneme-specific, they represent the corresponding viseme, thereby giving us good coverage of all visemes. Let I be a gray scale image. $p = [r, c]$ denotes the pair of integers r and c denoting the row and the column indices of the desired pixel p. Then $I(p) = I(r, c)$ denotes the pixel value at p. We represent the lip using a set of 22 points, $P = [p_1, p_2 ... p_{22}]$. P includes points of high gradient and where the contour of the lip has a high curvature. Such a set of points has been previously used by [25] and [26]. We manually annotate the 22 points in each training image. For a sample training image, the 22 points thus chosen are shown in Figure 3(a).

Since our region of interest is the lip, a rectangular patch of width $2w_{lip}$ and height $2h_{lip}$ is considered around the lip

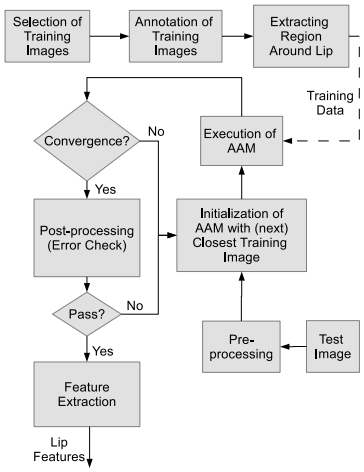
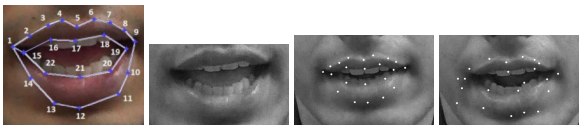


Figure 2: *An overview of the steps involved in computing lip features in PRAV Corpus.*



(a) (b) (c) (d)

Figure 3: *(a) Illustration of the 22 points representing the lip; (b) cropped area around the center of the lip; (c) and (d) show a comparison of a good and a bad AAM output. In (c), the AAM has converged to the boundaries of the lip, but in (d), the final contour has expanded to the chin.*

before extracting the lip features using AAM. The values for h_{lip} and w_{lip} are chosen to be large enough to capture the lip but small enough not to capture the nose, as seen in Figure 3(b). w_{lip} and h_{lip} are tuned separately for each subject to obtain the best performance using AAM (Table 1).

Initialization of the AAM: We observe that IC AAM often diverges within the rectangular area around the lip mainly due to the presence of the high gradient areas like the nose and the chin which attract the AAM contour. The AAM contour improves significantly when it is initialized better. For initialization, we detect the center of the lip and initialize the AAM contour around that point.

The center of the lip is determined by considering a window of size $2w_{sym}$ and $2h_{sym}$ (Table 1) around the horizontal and vertical axes around which the lower half of the face is maximally symmetric. We initialize the AAM with one of the annotated contours from the images in the training set whose lip image is most similar to the test lip image in the Euclidean sense.

Post-Processing: Sometimes, the contour obtained from the AAM is found to converge but does not take the shape of a lip. For example, Figures 3(c) and 3(d) show a good and a bad contour obtained from AAM respectively. To detect the incorrect contours, we run the contours from AAM through three errors checks, namely, ratio check, sequence check, and extrema check. These checks ensure that the points on the AAM contour do not provide abnormal shapes and maintain an order in both their row and column indices.

Table 1: *Parameter values used for all the subjects.*

Subject	w_{lip}	h_{lip}	w_{sym}	h_{sym}
M1	80	70	60	80
M2	140	90	70	140
F1	80	75	70	80
F2	100	65	40	100

Reinitialization: In the frames where AAM does not converge or does not pass the checks in the post-processing step, we rerun AAM by initializing with the annotated contour of the second closest training image to the test image. When the second initialization does not clear the checks, we initialize with the next closest training image and repeat till we exhaust all the images in the training set. This procedure results in a good final contour for a majority of the images. The images which do not make it through the checks even after all possible initializations constitute 1.63, 0.71, 2.24 and 0.44 percent of the total frames for subjects M1, M2, F1 and F2 respectively. We interpolate the contours for such missing frames. If three or more consecutive frames are missing, we annotate them manually.

Parameter Optimization In order to obtain the values of the parameters used for the error checks in post-processing, we use a development set comprising 10 randomly chosen sentences for each subject. This corresponds to 700 – 1000 frames depending on the subject. We experimentally optimize the parameters to detect most of the erroneous contours from AAM in the frames of the development set while at the same time reduce the false alarms.

Since the camera to subject distance was kept fixed during recording, determining w_{lip}, h_{lip}, w_{sym}, and h_{sym} did not require a large set of frames. These parameters are used to extract lip region from the face and with fixed camera-subject distance the effective facial area does not change significantly across different video recordings in the corpus. Thus, determining these parameters from few frames worked well for the entire corpus. However, the parameters used detecting errorneous contours in for post-processing need to be carefully selected since the quality of the AAM contour varies significantly depending on the lip shapes and phonetic context resulting in different degrees of errors in the contour estimated from AAM. Thus, a large development set is used to ensure that the selected parameters in post-processing are robust to lip shapes in different phonetic contexts.

Evaluation of the AAM: To evaluate the quality of the contour from AAM, we randomly choose a set of 15 frames for each subject and annotated them manually. Then we compute the contours from AAM which are compared with the manually annotated contours. We denote the manually annotated contour by P_{true} and the contour produced by AAM as P_{AAM}, each comprising 22 points denoted by $[r_j^{AAM}, c_j^{AAM}]$ and $[r_j^{true}, c_j^{true}]$ respectively, where $j = 1 \ldots 22$. We then compute the root mean squared error (RMSE)$= \sqrt{\frac{1}{22}\sum_{j=1}^{22}((r_j^{AAM} - r_j^{true})^2 + (c_j^{AAM} - c_j^{true})^2)}$. We report the mean and standard deviation of RMSE for each subject in Table 2. It is clear that the average RMSE of the contour estimated using AAM is of the order of a pixel. We thus conclude that the AAM with multiple initializations performs well in determining the lip contour, which is used to represent the lip features.

Table 2: *An evaluation of the performance of the AAM algorithm on test images*

Subject	M1	M2	F1	F2
RMSE(pixels)	0.8885	1.6822	1.2947	1.0898
SD(pixels)	0.1829	0.4147	0.3740	0.1617

4. Potential Utility of PRAV Corpus

Due to the richness in the phonetic contexts per subject, the PRAV Corpus could be useful for a number applications including audio-visual speech synthesis and audio-visual speech recognition. Some of these potential applications are briefly described below.

4.1. Audio-Visual Speech Synthesis

Modeling co-articulation is an important aspect in designing speech synthesizers. In audio-visual speech synthesis, this requires an availability of visemes in various phonetic contexts in the training data. In such scenarios, limited data is a bottleneck where the required phonetic context may not occur during the training phase [27]. Considering the phonetic richness of the PRAV Corpus, it would offer visemes in a wide variety of phonetic contexts. Moreover, visual speech synthesis is a task that inherently doesn't rely as much on subject diversity as it does on phonetic diversity. Even with a single subject and large training data, audio-visual speech can be synthesized for the respective subject, making PRAV Corpus ideal for audio-visual speech synthesis. This may be implemented as a text-to-speech synthesis speech systems using Avatars.

4.2. Audio-Visual Speech Recognition

The lip features provided with the PRAV Corpus could be readily used as visual features to develop models for visual speech recognition. The task of lip reading [28] could also benefit from having the lip features available from a large corpus such as PRAV. Moreover, instead of performing recognition with visual features alone, one can use them in tandem with features from the audio. For example, [29] demonstrated an improvement in the performance of the audio-visual speech recognition using such multi-modal features. Thus, the readily available visual features provided with the PRAV Corpus could be used to develop better audio-visual speech recognition models.

4.3. Audio-Visual Perception Studies

There have been several audio-visual perception studies to understand the role of audio and visual information and their interaction in sound perception. For example, the role of visual data as either an alternative to audio perception or as a complementary source of information in noisy conditions has been studied in the literature [30]. This role is quantified through the articulatory index (AI), a tool for estimating the relative importance of a particular frequency for speech intelligibility. Study by [31] demonstrates that visual features may not be as discriminative for plosive consonants as it is for the vowels. Thus, it could be interesting to estimate the performance of consonant and vowel recognition with and without the presence of visual data. Understanding the differences in the band-importance function for auditory and auditory-visual inputs has also explored by [32]. Since these studies have been performed on datasets having fewer phonemes, PRAV Corpus could aid such perceptual studies on account of the large number of sentences per subject along with its phonetic richness.

4.4. Comparison of Articulators from Visual Input and Acoustic to Articulatory Inversion

Acoustic to articulatory inversion (AAI) produces articulatory kinematics of a subject from a given speech signal. AAI is a difficult problem owing to the non-linear and one-to-many mapping between the acoustic and articulatory spaces. AAI is of two types: 1) subject-dependent [33], where the test subject's data is also available for training, 2) subject-independent [34], where the test subject's data is not available for training. The articulatory features from AAI have been shown to improve the speech recognition performance when used in addition to the acoustic features [35]. On the other hand, it is also known that the availability of the visual data such as speaker's facial video improves the speech recognition performance [36] particularly in noisy conditions [30]. The phonetic richness of the PRAV Corpus has been utilized to compare the utility of the AAI and visual features and to compute the complementarity of the information present in them with respect to the acoustic features [37].

4.5. Intelligibility and Naturalness of Facial Animation

The quality of a synthesized audio-visual speech depends not only on the intelligibility and naturalness of the speech but also on the naturalness of the facial expression including head movement and eye-blinking [38]. In particular, synthesis of intelligible consonants is a challenge where visual features play a significant role at low SNR [39]. Models for synthesizing videos with natural facial expressions could be learnt from PRAV Corpus which is rich with an individual's facial expression and gestures in a wide variety of phonetic contexts.

5. Conclusions and Future Work

We present a new audio-visual corpus, PRAV, which is phonetically rich and has the largest number of utterances per subject among all existing audio-visual corpora in the literature. In addition to the audio-visual data, we also provide lip features from the video and the routines to extract them. Due to its phonetic richness, the PRAV Corpus could be potentially used for a number of applications including, multi-modal speech processing, particularly synthesis, recognition and perception studies of audio-visual speech. We are planning to include additional features such as phonetic segmentation in the future versions of PRAV corpus. In order to increase the diversity of subjects in PRAV corpus, we plan to expand our dataset by including more age- and gender-balanced subjects. In these expansions, we also plan to include an affective component through a portrayal of emotions by the subjects.

6. Acknowledgement

We thank Department of Science and Technology, Government of India for their support in this work.

7. References

[1] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, and A. Mashari, "Audio visual speech recognition," IDIAP, Tech. Rep., 2000.

[2] S. A. Frisch and D. A. Nikjeh, "An audiovisual database of english speech sounds," *The Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2393–2394, 2003.

[3] S. Narayanan, E. Bresch, P. K. Ghosh, L. Goldstein, A. Katsamanis, Y. Kim, A. C. Lammert, M. I. Proctor, V. Ramanarayanan, Y. Zhu *et al.*, "A Multimodal Real-Time MRI Articulatory Corpus for Speech Research." in *INTERSPEECH*, 2011, pp. 837–840.

[4] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[5] C. Sanderson, "The VidTIMIT Database," IDIAP, Tech. Rep., 2002.

[6] C. McCool, S. Marcel, A. Hadid, M. Pietikainen, P. Matejka, J. Cernocky, N. Poh, J. Kittler, A. Larcher, C. Levy *et al.*, "Bimodal person recognition on a mobile phone: using mobile phone data," in *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2012, pp. 635–640.

[7] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty, "The Belfast induced natural emotion database," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 32–41, 2012.

[8] O. Onder, S. Zhalehpour, and C. E. Erdem, "A Turkish audio-visual emotional database," in *21st Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2013, pp. 1–4.

[9] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[10] I. Matthews, T. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, p. 2002, 2002.

[11] D. Schabus, M. Pucher, and G. Hofer, "Speaker-adaptive visual speech synthesis in the hmm-framework." in *INTERSPEECH*, 2012, pp. 979–982.

[12] S. Fu, R. Gutierrez-Osuna, A. Esposito, P. K. Kakumanu, and O. N. Garcia, "Audio/visual mapping with cross-modal hidden markov models," *Multimedia, IEEE Transactions on*, vol. 7, no. 2, pp. 243–252, 2005.

[13] W. Mattheyses, L. Latacz, and W. Verhelst, "Comprehensive many-to-many phoneme-to-viseme mapping and its application for concatenative visual speech synthesis," *Speech Communication*, vol. 55, no. 7, pp. 857–876, 2013.

[14] T. Ezzat and T. Poggio, "Visual speech synthesis by morphing visemes," *International Journal of Computer Vision*, vol. 38, no. 1, pp. 45–57, 2000.

[15] T. Choudhury, J. M. Rehg, V. Pavlović, and A. Pentland, "Boosting and structure learning in dynamic Bayesian networks for audio-visual speaker detection," in *Proceedings, 16th IEEE International Conference on Pattern Recognition*. IEEE, 2002, pp. 789–794.

[16] P. Phillips, P. Flynn, K. Bowyer, R. Bruegge, P. Grother, G. Quinn, and M. Pruitt, "Distinguishing identical twins by face recognition," in *IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG)*, March 2011, pp. 185–192.

[17] K. Messer, J. Matas, J. Kittler, and K. Jonsson, "XM2VTSDB: The Extended M2VTS Database," in *In Second International Conference on Audio and Video-based Biometric Person Authentication*, 1999, pp. 72–77.

[18] I. Anina, Z. Zhou, G. Zhao, and M. Pietikäinen, "OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis," in *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 1. IEEE, 2015, pp. 1–5.

[19] D. Burnham, D. Estival, S. Fazio, J. Viethen, F. Cox, R. Dale, S. Cassidy, J. Epps, R. Togneri, M. Wagner *et al.*, "Building an Audio-Visual Corpus of Australian English: Large Corpus Collection with an Economical Portable and Replicable Black Box." in *INTERSPEECH*, 2011, pp. 841–844.

[20] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, p. 27403, 1993.

[21] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," 2001.

[22] A. D. Team, "Audacity (version 1.2. 6) [computer software]," *Available: audacity. sourceforge. net/download*, 2008.

[23] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pp. 681–685, 2001.

[24] L. Vezzaro, "ICAAM - Inverse compositional Active Appearance Models," 2011. [Online]. Available: http://in.mathworks.com/matlabcentral/fileexchange/32704-icaam
-inverse-compositional-active-appearance-models

[25] C. Bregler and Y. Konig, "Eigenlips for robust speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2. IEEE, 1994, pp. II–669.

[26] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, and A. Mashari, "Audio visual speech recognition," IDIAP, Tech. Rep., 2000.

[27] W. Mattheyses and W. Verhelst, "Audiovisual speech synthesis: An overview of the state-of-the-art," *Speech Communication*, vol. 66, pp. 182–217, 2015.

[28] G. Zhao, M. Barnard, and M. Pietikäinen, "Lipreading with local spatiotemporal descriptors," *Multimedia, IEEE Transactions on*, vol. 11, no. 7, pp. 1254–1265, 2009.

[29] G. Potamianos and C. Neti, "Audio-visual speech recognition in challenging environments." in *INTERSPEECH*, 2003, pp. 1293–1296.

[30] J. MacDonald and H. McGurk, "Visual influences on speech perception processes," *Perception & Psychophysics*, vol. 24, no. 3, pp. 253–257, 1978.

[31] D. W. Massaro and M. M. Cohen, "Evaluation and integration of visual and auditory information in speech perception." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 9, no. 5, pp. 753–771, 1983.

[32] K. W. Grant and L. D. Braida, "Evaluating the articulation index for auditory–visual input," *The Journal of the Acoustical Society of America*, vol. 89, no. 6, pp. 2952–2960, 1991.

[33] S. Hiroya and M. Honda, "Speaker adaptation method for acoustic-to-articulatory inversion using an HMM-based speech production model," *IEICE TRANSACTIONS on Information and Systems*, vol. 87, no. 5, pp. 1071–1078, 2004.

[34] A. Afshan and P. K. Ghosh, "Improved subject-independent acoustic-to-articulatory inversion," *Speech Communication*, vol. 66, pp. 1–16, 2015.

[35] J. Frankel, "Linear dynamic models for automatic speech recognition," *Doctoral Thesis at The University of Edinburgh. College of Science and Engineering. School of Informatics*, 2004.

[36] A. MacLeod and Q. Summerfield, "Quantifying the contribution of vision to speech perception in noise," *British journal of audiology*, vol. 21, no. 2, pp. 131–141, 1987.

[37] A. Narwekar and P. K. Ghosh, "A comparative study of articulatory features from facial video and acoustic-to-articulatory inversion for phonetic discrimination," in *IEEE International Conference on Signal Processing and Communications (SPCOM)*, 2016, pp. 1–5.

[38] Z. Deng, J. P. Lewis, and U. Neumann, "Automated eye motion using texture synthesis," *Computer Graphics and Applications, IEEE*, vol. 25, no. 2, pp. 24–30, 2005.

[39] C. A. Binnie, A. A. Montgomery, and P. L. Jackson, "Auditory and visual contributions to the perception of consonants," *Journal of speech, language, and hearing research*, vol. 17, no. 4, pp. 619–630, 1974.