International Journal of Speech Technology
https://doi.org/10.1007/s10772-019-09655-0

Viseme set identification from Malayalam phonemes and allophones

K. T. Bibish Kumar[1] · R. K. Sunil Kumar[2] · E. P. A. Sandesh[3] · S. Sourabh[1] · V. L. Lajish[3]

Received: 31 March 2019 / Accepted: 24 October 2019

Abstract

Knowledge about phoneme and viseme in a language is a vital component in the making of any speech-based applications in that language. A phoneme is an atomic unit in an acoustic speech that can differentiate meaning. Viseme is the equivalent atomic unit in the visual realm which describes distinct dynamic visual speech gestures. The initial phase of the paper introduces a many-to-one phoneme-to-viseme mapping for the Malayalam language based on linguistic knowledge and data-driven approach. At the next stage, the coarticulation effect in the visual speech studied by creating many-to-many allophone-to-viseme mapping based on the data-driven approach only. Since the linguistic history in the visual realm was less explored in the Malayalam language, both mapping methods make use of K-mean data clustering algorithm. The optimum cluster determined by using the Gap statistic method with prior knowledge about the range of clusters. This work was carried out on Malayalam audio-visual speech database created by the authors of this paper with consist of 50 isolated phonemes and 106 connected words. From 50 isolated Malayalam phonemes, 14 viseme were linguistically identified and compared with results obtained from a data-driven approach as whole phonemes and consonant phonemes. The many-to-many mapping studied as a whole allophone, vowel allophones, and consonant allophones. Geometric and DCT based parameters are extracted and examined to find the parametric phoneme and allophone clustering in the visual domain.

Keywords Phonemes · Allophones · Visemes · K-mean clustering · Gap statistic

✉ K. T. Bibish Kumar
bibishkrishna@gmail.com

R. K. Sunil Kumar
seuron74@gmail.com

E. P. A. Sandesh
sandeshepa@gmail.com

S. Sourabh
sourabhsuresh5@gmail.com

V. L. Lajish
lajishvl@gmail.com

[1] Computer Speech & Intelligence Research Centre, Department of Physics, Govt. College, Madappally, Vadakara, Calicut, Kerala 673102, India

[2] School of Information Science and Technology, Kannur University, Kannur, India

[3] Department of Computer Science, University of Calicut, Calicut, Kerala 673 635, India

1 Introduction

Speech is bimodal; that is the most frequent communication system between humans and involves the understanding of the auditory and visual channels. The contribution of the visual part in the judgment of speech, especially in a noisy environment, is a fact. The visible organs of the articulatory system of human speech production consist of upper and lower lips, teeth, tongue, and lower jaw. The lips, tongue, and jaw will be the actively visible articulators used in language production. Analyzing the most dynamic active visible articulator, the lips, is the most crucial component in the visual speech analytics framework for recognition and synthesis.

Phonemes in a speech would be the nuclear sound units necessary to symbolize all words in that speech. On the other hand, the visual equivalent of a phoneme has several features which require a comprehensive study of this phoneme-to-viseme mapping region. For many years of study in visual language, it has gained considerable alterations in its definition. A viseme can contemplate regarding articulatory gestures such as mouth opening, teeth, and tongue vulnerability

Published online: 04 November 2019

that have to generate different phoneme as in Fisher (1968). An equivalent definition which has been used extensively in literature is a viseme for a set of phonemes which has a similar visual look like in Bear and Harvey (2016) and Bozkurt et al. (2007). The static viseme does not account for the coarticulation effect of a visual address. On the other hand, the current description of viseme is a lively visual language unit that describes distinct speech movements of the visual speech articulators as in Taylor et al. (2012).

Analysis of the visual speech signal and extracting the viseme set shows satisfactory improvement in recognition of language component. A mapping between phoneme and viseme needs to be anticipated to synchronize the mouth form of distinct sounds. Before addressing the problems related to speaker variability (Bear and Harvey 2018; Farooq et al. 2015), pose (Lucey and Potamianos 2007), choice of classifier technology (Noda et al. 2015; Sarma and Sarma 2015) and recording device (Blokland and Anderson 1998; Saitoh and Konishi 2010), the primary task in visual speech analysis to decode the visual information from the lips. Since the extent of deformation of lips restricts as a result of facial muscles compared to strain of the vocal organs, the viseme set in a speech is always smaller than the phoneme set. For developing the viseme set, the literature suggests two different approaches: linguistic and data-driven approach (Mattheyses et al. Mattheyses et al. 2013; Jachimski et al. 2018). According to linguistic understanding, phonemes with the similar visual appearance of active articulators treat as a viseme. In the data-driven strategy, the visual speech analyzed in the aspect by extracting essential features from the lip area and group the features based on the similarity measurement. The linguistic strategy is highly dependent on perception ability of the linguistic or trained individual, which accurately represent the human lip-reading nature and a time-consuming process. The data-driven approach provides a less time-consuming endeavour, but computational analysis of visual address profoundly depends on the choice of visual characteristic, which can be language-dependent. In human understanding, a holistic perspective is much more important than parts (Gestalt perception theory) but in computer vision parts (pixels) is much more significant than the whole (picture) (Alexandre and Tavares 2010). As a result of this battle, the data-driven approach alone cannot be able to mimic human perception accurately, and linguistic knowledge alone cannot have the ability to examine substantial visual speech information. Therefore, a linguistic involved data-driven approach can make valuable of individual perception modelling from linguistic approach and the computational easiness out of a data-driven approach.

Aim of this research work is to identify the viseme set in Malayalam language using linguistic involved data-driven approach by consciously neglecting the problems in visual speech as discussed above. One of the haunting problems in visual speech analysis is the availability of phonetically rich database in the concerned language. Besides, there is no collective agreement in different aspects like how large enough it should be, the number of speakers and the facial variability required to generalize the whole population for better improvement in the computational output. As an initial work in Malayalam visual speech analysis area, we recorded a multimodal speech database of 23 native speakers of Kerala by capturing the lip region of the speaker's face. Malayalam is a syllable based language written with a syllabic alphabet in which all consonants have an inherent vowel/a/. The database includes vowel and consonant–vowel syllable which altogether comprises of 50 phonemes and 106 words which capture all contextual variations of 50 phonemes namely the allophone.

The crucial step involved in the visual speech analysis is the identification of relevant static frame from the recorded video, which contains the visual appearance of the underlying phoneme. Based on the linguistic mapping, frames having similar visual appearance were selected manually for each speaker, which will minimize the error rate during data-driven analysis. The main problem to be addressed in the data-driven approach is which visual feature has to selected to model the human perception. Only by considering the visual speech properties of the underlined language can solve it. For uttering Malayalam sound, the tongue plays a vital role in terms of speed and flexibility, which makes distinct from other languages. The degree of presence of teeth, and oral cavity and the shape of the lip can be modelled using geometric features of lips and deformation in the appearance of lips and tongue can be modelled through Discrete Cosine Transform (DCT) feature. So in this work, the mathematical analysis is initiated by extracting the Geometric features and DCT features from the selected frames.

The next question arises how to identify the viseme set from visual features and what should be the number of viseme. Researchers conducted a wide range of techniques to establish the mapping between phoneme and viseme, but there are no reliable and unambiguous methods to confirm that this is better than the other. In this work, we categorized the visual similar feature vectors to viseme groups by combining the K-Mean clustering method (Miglani and Garg 2013) and Gap statistic method (Mohajer et al. 2011). Since K-Mean requires a pre-determined cluster number, the gap statistic method identifies the optimal cluster number from a range of clusters by exploring the language knowledge. The correlation between static visual speech unit and phoneme is carried out by developing a many-to-one phoneme-to-viseme mapping from isolated phonemes based on linguistic knowledge and by clustering in the parametric space using different visual features. The coarticulation effect in visual speech accounted by developing many-to-many allophone to viseme mapping using a data-driven approach alone.

Vowel-only and consonant only mapping is also mentioned in this work.

The content of this paper arranged as follows: The concept of viseme with viseme formation strategies explained in Sect. 2. Section 3 explains the audio-visual Malayalam speech database developed for this work. Section 4 explains the different phoneme to viseme mappings derived using linguistic and parametric approaches. Section 5 discusses the data-driven allophone to viseme mapping performed to obtain a viseme set which incorporates coarticulation effect. Section 6 concludes the chapter with future directions.

2 Viseme set formation approaches and known viseme sets

Many researchers have analysed the importance of the phoneme to viseme mapping. The phonemes which have almost the same visual mouth appearance grouped to a single viseme class. In literature, many mappings reported (Bear et al. 2017), range of visemes in a language varies between 10 and 20. The number and nature of viseme are language-dependent. Hence a language-specific exploration is needed for establishing the viseme set for a particular language. Traditionally there are three approaches for obtaining visemes from a many to one mapping: linguistic knowledge-based (Aschenberner and Weiss 2005; Meier et al. 2000; Setyati et al. 2015), perception experiments with human subjects (Fisher 1968; Lalitha and Thyagharajan 2016; Montgomery and Jackson 1983) and data-driven approach (Damien et al. 2009; Hazen et al. 2004; Mattheyses et al. 2013; Melenchón et al. 2007). Some authors blend linguistic and perception experiments based on approaches and name them as subjective assessments. Using a subjective approach, viseme classes are defined through linguistic knowledge and prediction of phonemes having a similar visual appearance. Viseme classes created by clustering of phonemes based on features extracted from the mouth region is the highlight of a data-driven approach. Most of the work has reported in European languages, but in India, only a few such as Hindi (Mishra et al. 2013; Upadhyaya et al. 2015; Varshney et al. 2014) and Marathi (Brahme and Bhadade 2017) have been studied in visual speech. Mishra et al. (2013) made a Hindi phoneme speech recognition system using DCT as a visual feature and MFCC (Mel Frequency Cepstral Coefficient) as an audio feature which reports better recognition in a noisy environment. Upadhyaya et al. (2015) Studied the performance of audiovisual speech recognition system under diverse noisy audio condition using a different combination of image-based features. They studied the dependency of recognition rate on a different type of visual features and nature of the acoustic noise used (Varshney et al. 2014). Clustered 23 phonemes into five viseme classes based data-driven approach using DCT as a visual feature. They also have done the viseme recognition task by integrating with audio features and reported improvement in the recognition rate. Brahme and Bhadade (2017) Presented phoneme viseme mapping for Marathi language based on linguistic approach alone. They derived 13 viseme classes, including silence from 44 phonemes. The center of discussion so far shows attempts have not yet reached the realm of successful development of visual speech technology in the Indian language. This research work is going to be the initial study in Malayalam phoneme to viseme mapping based on linguistic assessments and data-driven approach and allophone-to-viseme mapping based on a data-driven approach alone. Table 1 summarizes a brief background of established viseme mapping in terms of language, methodology and some special features.

Almost all viseme maps have developed either based on linguistic approach or data-driven approach or both. Besides, only a few viseme maps have studied for the coarticulation effects of visual speech, which might be the lack of database containing all contextual variations in the language concerned. This work uses both approaches and also studied the contextual variation of phonemes by developing allophone-to-viseme mapping, which explained in the proceeding sections.

3 Malayalam audio-visual speech database and data acquisition

Malayalam is an Indian language spoken by 40 million people as an official language in Kerala, and the union territories of Lakshadweep and Puducherry. It is the youngest in the Dravidian language family and conferred the classical language status by the Government of India in 2013. Even though Malayalam is developed double rooted from Sanskrit and Tamil but display wide variation in another aspect, which makes Malayalam a distinctive in Indian languages.

In this work, an audio-visual Malayalam speech database is created from 23 native speakers of Kerala (18 females and five males) by capturing the lip region of the speaker's face in normal lighting conditions. The utterances are taken in the form of 'silence-phoneme/word-silence' fashion. The language material of this database contains ten vowel phonemes, two diphthongs, and 38 consonant–vowel syllable which altogether comprises of 50 phonemes and 106 words which capture all contextual variations of 50 phonemes which include 75 consonant allophones, 28 vowel allophones and three allophones corresponding to diphthongs as in http://www.cmltemu.in/phonetic/#/. Tables 2 and 3 shows the list of Malayalam vowel phonemes and consonant phonemes with its corresponding allophones. The phonemes are arranged in a front to back manner (from lip to glottal) of articulator's position.

Table 1 Summary of known viseme sets

Author (year)	Linguistic information	Implementing method	Special features
Fisher (1968)	23 initial consonants—5 viseme classes 20 final consonants—5 viseme classes	Multiple-choice intelligibility test Confusion matrix	Studied visual perception of initial and final consonants Mapping is done subjectively Fisher introduced the term viseme which is a compound word of visual and phoneme
Franks (1972)	Initial consonants		Consonants are only studied
Binnie (1976)	20 English consonants—9 viseme classes	Human testing Confusion matrix	Consonants are only studied 20 English consonants were combined with the vowel/a/to form 20 CV syllables 34 female observers have participated in the testing process Mapping is done subjectively
Janet Jeffers (1971)	American English 43 phonemes—11 viseme classes	Pure linguistic approach	
Montgomery and Jackson (1983)	American English 15 vowels and diphthongs	3 methods—perceptual analysis using confusion matrix, physical measurements (height, width, area, acoustical and visual duration) and correlation between the two	The study was restricted to vowels only
Lander (1999)	35 phonemes—12 classic Disney mouth position	Linguistic approach	Facial animation
Neti et al. (2000)	42 phonemes—12 viseme classes (excluding silence)	A mixture of a linguistic and data-driven approach Decision tree-based HMM state clustering method Models are trained using DCT visual features	IBM ViaVoice database was used
Lee and Yook (2002)	41 phonemes—14 viseme classes 7 vowel, 6 consonant and 1 silence viseme	Assumed to be a linguistic approach HMM, modelling Used context-independent recognition units (phone model) Produced a sequence of viseme symbols from speech waveform	TIMIT speech database Two approaches in building viseme recognizer-viseme HMMs and phoneme HMMs.
Hazen et al. (2004)	American English 50 phonemes—14 viseme classes There are 54 phonemes but 4 phonemes were merged to get 50 phonemes	Data-driven approach Agglomerative hierarchical clustering algorithm Bottom-up clustering using maximum Bhattacharyya distances 96-dimension stacked PCA feature vectors	AV-TIMIT speech database Viseme was represented by three consecutive frames with middle frame describe the static viseme of each phoneme Before clustering, some phonemes were merged
Aschenberner and Weiss (2005)	German language 42 phonemes—15 viseme classes	Linguistic approach	Applied for speech synthesis
Bozkurt et al. (2007)	American English 46 phonemes—16 viseme classes	Linguistics approach HMM, the model was used for lip animation with audio speech as input Contextual information was implemented using tri-phone model	TIMIT speech database Compared phone, tri-phone, viseme and tri-viseme based HMM structure for lip animation Acoustic observation consists of 12 MFCC, energy, delta and acceleration coefficients resulting in 39 feature-length Applied for visual speech synthesis

Table 1 (continued)

Author (year)	Linguistic information	Implementing method	Special features
Melenchón et al. (2007)	Spanish language 12 allophones—6 viseme classes	Data-driven approach 12 PCA coefficients were used as a feature vector	Three speakers utter 12 Spanish sentences
Chitu and Rothkrantz (2009)	Dutch language 40 phonemes—18 viseme classes	Confusion matrix	
Damien et al. (2009)	Arabic language 28 phonemes—10 viseme classes	Data-driven approach Geometrical features used	Four speakers utter four types of word sequences
Yu et al. (2010)	50 words—60 classes of visual speech units (VSU)	Data-driven approach Used Expectation–Maximization Principal Component Analysis (EM-PCA) as a feature extraction method Based on HMM classification	Introduced new term "Visual Speech Unit (VSU)" which include transition information between consecutive visemes Two speakers utter a total of 50 words
Mattheyses et al. (2013)	Dutch language Many-to-many phoneme-to- viseme mapping	Data-driven approach AAM (Active Appearance Model)-based representation of mouth region Tree-based and k-means clustering approach was used	Coarticulation effect was studied Applied for visual speech synthesis
Seko et al. (2013)	Japanese language 40 phonemes—14 viseme classes (excluding silence)	HMM modelling	CENSREC-1-AV database was used
Aghaahmadi et al. (2013)	Persian language 23 consonant phonemes—7 classes	Data-driven approach—Eigen vectors as feature Proposed a new method to cluster the dataset A subjective test is also used	The proposed method is tested on an American English database
Taylor et al. (2015)	Created many-to-many mapping Approximately 50,000 visual speech gestures—150 dynamic viseme classes	Clustered the speech gestures identified by AAM (Active Appearance Model) of jaw and lips 20 Dimension feature vector entirely describes the shape and appearance information Dynamic visemes were learned entirely from visual data	KB-2K database was used A single actor recites 2542 phonetically balanced sentences from the TIMIT database Applied for automatic redubbing of video
Setyati et al. (2015)	Indonesian language 49 phonemes—12 viseme classes	Linguistic approach	Used blend shape models for analysing the facial images Ten speakers were used for this study
Bear et al. (2017)	46 phonemes—visemes ranging from 2 to 45	Viseme classes were obtained based upon the mapping of articulated phonemes, which was confused during phoneme recognition, into viseme groups	Designed speaker-dependent viseme classes Studied on LiLIR dataset 12 British speakers utter about 1000 words totally

Table 2 Malayalam vowels and diphthongs with its allophonic variations

SI. no.	Vowel phoneme IPA	Vowel allophone IPA	SI. no.	Vowel phoneme IPA	Vowel allophone IPA
1	ഇ /i/	[i] [^{y}i] [y^{i}]	7	ഉ /u/	[^{w}u] [u^{w}] [ɯ] [ə] [ə*] [ɯ^{v}] [U]
2	ഈ /i:/	[^{y}i:] [i:]	8	ഊ /u:/	[^{w}u:] [u]
3	എ /e/	[^{y}e] [e^{y}] [E]	9	ഒ /o/	[^{w}O] [O]
4	ഏ /e:/	[^{y}e:] [e^{r}:] [e:]	10	ഓ /o:/	[^{w}O:] [O]
5	അ /a/	[ʌ] [A]	11	ഐ /ai/	[ai] [ei]
6	ആ /a:/	[a:] [a]	12	ഔ-/au/	[au]

Table 3 Malayalam consonants with its allophonic variations

SI. no.	Consonant phoneme IPA	Consonant allophone IPA	SI. no.	Consonant phoneme IPA	Consonant allophone IPA	SI. no.	Consonant phoneme IPA	Consonant allophone IPA
1	പ് /P/	[p] [β] [b] [P]	14	സ് /s/	[s]	27	ഛ് /c^{h}/	[c^{h}] [C^{h}]
2	ഫ് /p^{h}/	[p^{h}]	15	ര് /r/	[r]	28	ജ് /ɟ/	[J] [j]
3	ബ് /b/	[B] [b]	16	റ് /ṛ/	[ṛ]	29	ഝ് /ɟ^{h}/	[ɟ^{h}]
4	ഭ് /b^{h} /	[b^{h}]	17	ല് /ൽ /l/	[l]	30	ഞ് /ɲ/	[ɲ]
5	മ് /m/	[m̠^{h}] [M] [m]	18	ട് /ʈ/	[ɖ] [ɽ] [ʈ] [T]	31	ശ് /ʃ/	[ʃ]
6	വ് /v/	[w] [v]	19	ഠ് /ʈ^{h}/	[ʈ^{h}] [T^{h}]	32	യ് /y/	[y]
7	ത് /t/	[t] [t''] [ð] [d̠]	20	ഡ് /ɖ/	[ɖ]	33	ക് /k/	[k] [kj] [ɣ̊] [ɠ] [ʈ] [K]
8	ഥ് /t^{h}/	[t^{h}]	21	ഢ് /ɖ^{h}/	[ɖ^{h}]	34	ഖ് /k^{h}/	[k^{h}] [K^{h}] [K^{h}]
9	ദ് /d/	[ḍ] [d]	22	ണ് /ɳ/	[ɳ]	35	ഗ് /g/	[G] [g]
10	ധ് /d^{h}/	[d^{h}]	23	ഷ് /ʂ/	[ʂ]	36	ഘ് /g^{h}/	[g^{h}]
11	ന് /n̪/	[n̠] [n]	24	ള് /ൾ/ɭ/	[ɭ]	37	ങ് /ŋ/	[ŋ] [ŋj] [ŋ<] [ŋ>] [ŋ']
12	റ്റ് /r̠/	[d] [t]	25	ഴ് /ʐ/	[ʐ]			
13	ന് /n/	[n^{h}] [n]	26	ച് /c/	[c] [c̬] [ɟ] [C]	38	ഹ് /h/	[H] [h]

A high-quality visual speech is recorded from the speaker's mouth region with a resolution of 1280 × 720 having a frame rate of 25 fps in MP4 format. For an individual speaker, the approximate footage span is 5 min for all isolated phonemes and 20 min for connected words. After documenting the multimodal speech database, the visual speech mode alone is further examined for viseme mapping. In future, the audio speech mode can also be analysed along with the visual mode for recognition and synthesis task by measuring the audio-visual speech synchrony. The crucial step involved in mapping procedure is to the identification of relevant static frame from the recorded video, which contains the visual look of the underlying phoneme. This work is performed by rending the expertise from the linguistic peoples. In Malayalam, the language component in the audio speech, i.e., phoneme, is linguistically categorised according to articulation points and manners as in http://www.cmltemu.in/phonetic/#/. This is the primary work to classify visual speech based on a linguistic basis. Linguistic people select the relevant frame based on the articulatory rules and their expertise to capture the dynamical variability of the speaker's mouth appearance. Meanwhile, the first author of this work got training from the linguistic people for selecting the relevant frame. After selecting the frame for all phonemes of two speakers by the linguistic people, this work is carried out to other speakers by the trained individual. The selected frame also undergoes a post-selection endeavour, that is performed by the rest of the authors of this work, to minimise the individual biasing and error rate during further analysis. To encode the temporal variant in the visual speech of the underlined phoneme, two frames in the left side and right side of the chosen frame are also selected. Figure 1 shows the sequential arrangement of the frame which will capture the visual dynamics of underline phonemes.

The chosen frame comprises insignificant area/regions like background screen, hairs, ornaments, etc., which does not contribute anything to the visual speech has the potential to induce serious influence in the analysis section. Before analyzing the linguistic and data-driven domain, undesirable areas ought to be removed in the frames. Manual cropping cannot have the capability to take care of the issues linked to the uniform framework area required to catch the dynamical variability in the appearance of underlying phoneme and speaker variability. In this work, we embraced a semi-automatic cropping method to deal with the issue mentioned above and can further use as a by-product in the data-driven approach. The very first step in semi-automatic cropping way is manually extracting the lip shape information since the dynamicity related to the appearance of phoneme and speaker is embedded in the lip shape. To extract the lip shape information, accurate lip contour is needed. Lip contour is obtained by linearly curve fitting the manually marked landmark points on each image. Lip contour is defined using 36 shape feature points on the lip. Twenty landmark points represent the outer lip contour, and 16 landmark points are used to describe the inner lip contour. Each landmark point contains x-coordinate and y-coordinate, that is the pixel position in the image. Figure 2 shows the images with lip contour marked manually using 36 landmark points. The second step in the semi-automatic cropping method is automatically to crop the frame using these landmark points. For this, the centroid of lip contour is estimated from x and y coordinates of the landmark points using the equation for the centroid of a finite set of points. After this, a frame of dimension 600 × 500 is cropped whose mathematical centre passes through the centroid of the lip shape of the original

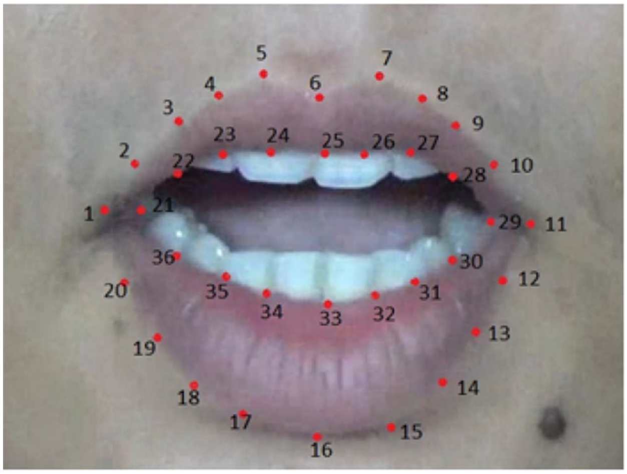


Fig. 2 Manual labelling landmark points in ROIs

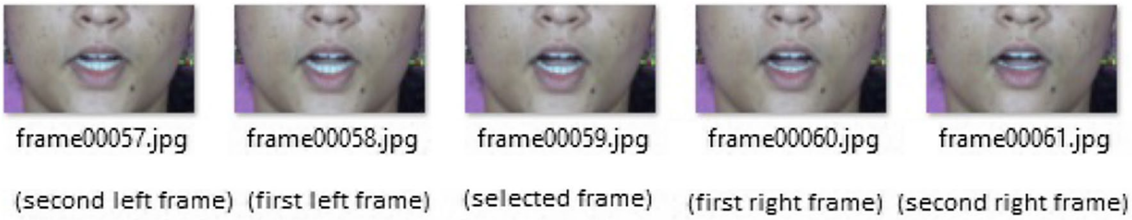


Fig. 1 Sequential frames for phoneme—അ /a/

frame. The dimension of the frame is iteratively chosen to embed the lip region only.

4 Malayalam phoneme-to-viseme/many-to-one mapping

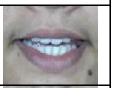
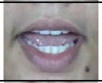
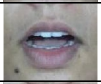
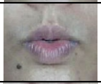
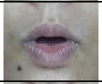
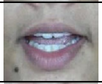
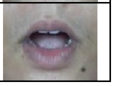
In this work, linguistic and data-driven approaches adopted for discovering the viseme set. The linguistic approach is carried out under the guidance of linguistic persons by linguistically analysing the utterance style of the speakers. In the data-driven approach, the mathematical representation of visual speech extracted and cluster based on the similarity measurement. Literature Ahmad et al. (2008), Stewart et al. (2008), and Sui et al. (2016) shows a wide range of visual features based on the type of information embedded in it, which are geometric-based, image-based and model-based. Geometric-based features explicitly simulate the measurement of mouth concerning height, weight, area, perimeter etc. by analysing the pixels in the lip boundary. Centroid distance and Fourier transcriptor (Websdale and Milner 2015) belong to this category. Image-based feature considers all pixels in the region of interest (ROI) are informative to represent the speech. In this method, the ROI is transformed into a different domain, thereby capturing the most informative components. Discrete Cosine Transform (DCT) (Farooq et al. 2013; McLaren and Lei 2015; Puviarasan and Palanivel 2011), Discrete Wavelet Transform (DWT) (Morade 2016; Morade and Patnaik 2014), Principal Component Analysis (PCA) (Rajavel and Sathidevi 2009; Xiaopeng et al. 2006) and Linear Discriminant Analysis (LDA) (Alizadeh et al. 2008) and the combinations (He and Zhang 2009) belong to this category. Model-based features create a mathematical model to extract visual information with high computational complexity. Active Shape Model (ASM) and Active Appearance Model (AAM) (Baswaraj et al. 2012; Biswas et al. 2015) belongs to this class. On account of this diversity from the lip movements of the speaker from the world, just language mining can figure out this matter. Due to the diversity in the lip movement of the speaker's in the world, only language exploration can solve this issue. For uttering Malayalam sound, the tongue plays a critical role concerning flexibility and speed, which makes distinct from other languages. The amount of existence of teeth, and oral cavity and the shape of the lip can be modelled with geometric features of lips and deformation in the appearance of lips and tongue can be modelled through Discrete Cosine Transform (DCT) feature. The visual speech attributes are then clustered to identify the visual equivalent of the phoneme. Clustering is a vital step in data mining to discover the hidden pattern of an unlabelled dataset based on mathematical measurement. This method divides the dataset into smaller subclasses that have high intra-class similarity and low inter-class similarity. There are two widely used clustering algorithms such as Hierarchical (Agglomerative and Divisive) (Li et al. 2014; Madhulatha 2012), Partitional (K-mean) (Jain 2010; Miglani and Garg 2013). They explore the partition of data objects based on the number of clusters. On the other hand, the number of clusters obtained from such approaches is highly sensitive to the nature of the dataset. Thus, identifying the optimal number of cluster is a significant endeavour, and can be carried out using the Gap statistic method (Mohajer et al. 2011; Tibshirani et al. 2001). To cluster large and highly correlated dataset, K-mean clustering, together with the Gap statistic method, is employed for optimum cluster selection in this work. To our knowledge, this is the very first work that employs the gap statistic method in the development of the viseme map.

4.1 Linguistic approach

In Malayalam, the language component in the audio speech, i.e., phoneme, is linguistically categorized according to articulation points and manners as in http://www.cmltemu.in/phonetic/#/. The visual speech appearance depends primarily on the lip and lower jaw movements. Visibility of teeth and tongue is also a leading element. While uttering vowel phonemes, most of the sound characters appear in the shape of the lip by a wide-open and outward posture of lips. However, consonant phonemes are produced by touching the active articulator tongue at different places inside the mouth area whose dynamics is not visible. Thus the extent of appearance of active articulators is the crucial factor to characterize the consonant phonemes. This section explores the possibilities of forming a viseme set from linguistic knowledge about the language and its phoneme set by rending the expertise from the linguistic peoples.

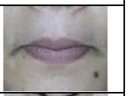
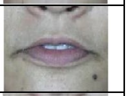
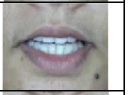
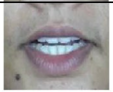
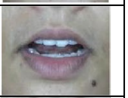
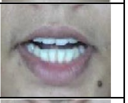
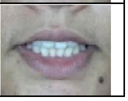
The vowel sound is generated by the flow of air from the larynx to the lips with no obstruction in the mouth region. Linguistically the five short vowel phonemes are distinguished by the position of the tongue as front-high, front-mid, central–low, back-high and back-low. Since these tongue positions are partially visible, along with the shape information is utilized to categorize the visual equivalent of phonemes. The impact of time in long vowels is weak to differentiate visually from corresponding short vowel phonemes. The front-high vowel ഇ /i/ exhibits a wide horizontal opening and vertical opening by central-low vowel അ /a/. Front-mid vowel എ /e/ visually placed between front-high and central-low vowel. Both back-high ഉ /u/ and back-mid ഒ /o/ vowel shows a rounded lip shape with a discriminating outward posture for the back-high vowel. The quick gliding of tongue from one vowel to another characterizes a diphthong. The diphthong ഐ /ai/ is made from the transition from അ /a/ to ഇ /i/ and ഔ /au/ is obtained from അ /a/ to ഉ /u/. Since the visual characterization of vowels is taken from the selected frame and visual signature of the diphthong captured from the temporal variation of the selected frame. Due to the diversity, two diphthongs are assigned to separate viseme classes. In short, each vowel assigned to a separate viseme classes, but Monophthong short and long phonemes of the same vowel placed in the same class. The viseme set for vowels and diphthongs in Malayalam formed from linguistic understanding is given in Table 4.

Table 4 Linguistic classification of monophthongal and diphthongal vowel phonemes

Viseme	Viseme class	Phoneme with IPA	Viseme in frame
1	Front, High—Vowel	ഇ /i/ , ഈ /i:/	
2	Front, Mid—Vowel	എ /e/ , ഏ /e:/	
3	Central, Low —Vowel	അ /a/ , ആ /a:/	
4	Back, High—Vowel	ഉ /u/ , ഊ /u:/	
5	Back, Mid—Vowel	ഒ /o/ , ഓ /o:/	
6	Diphthong 1	ഐ /ai/	
7	Diphthong 2	ഔ /au/	

In contrary to the vowel phoneme/sound, the consonant sound articulated with the complete or partial closure of the vocal tract, which is visually distinguishable only by considering the lip appearance. Most visually distinctive sound element in consonant class is bilabial sound. While uttering this sound, the lips are kept closed with a slight strain in the facial muscles. The first consonant viseme class formed from bilabial plosives expect ഫ് /ph/. വ-/va/, the only true labiodentals in the Malayalam and the Bilabial—Plosive-voiceless aspirated ഫ് /ph/ which is visually different from other Bilabial phonemes placed in the next viseme class due to the feeble presence of teeth. Viseme 10 consist of dental consonants, which has got the maximum teeth visibility and some traces of tongue tip. The velar consonants and the only glottal phoneme ഹ-ha placed in the next viseme class since their place of articulation is back of the tongue and has the same visual appearance. The tongue is further backward in alveolar consonants which are less visible and grouped in viseme class 12. Retroflex consonants are produced by curling the tongue backwardly and touches the front part of the hard palate which produces the same visual appearance and thereby assigning it as new viseme group. Viseme 14 linguistically characterized as palatal consonants which produced by touching the tongue towards the hard palate. In brief, 50 isolated Malayalam phonemes mapped into 14 viseme classes (Table 5).

Table 5 Linguistic classification of consonant phonemes

Viseme	Viseme class	Phoneme with IPA	Viseme in frame
8	Bilabial— Plosive-voiced and voiceless unaspirated, Nasal	പ് /p/ , ബ് /b/ , ഭ് /bh/, മ് /m/	
9	Bilabial— Plosive-voiceless aspirated And Labiodental	ഫ് /ph/, വ് /v/	
10	Dental	ത് /t/ , ഥ് /th/ , ദ് /d/ , ധ് /dh/ , ന് /n̪/	
11	Velar Glottal	ക് /k/ , ഖ് /kh/ , ഗ് /g/ , ഘ് /gh/ , ങ് /ŋ/ ഹ് /h/	
12	Alveolar	റ്റ് /r̠/ , ന് /n/ , സ് /s/ , ര് /r/ , റ് /ṛ/ , ല് /l/	
13	Retroflex	ട് /ʈ/ , ഠ് /ʈh/ , ഡ് /ɖ/ , ഢ് /ɖh/ , ണ് /ɳ/ , ഷ് /ʂ/ , ള് /ɭ/, ഴ് /ʐ/	
14	Palatal	ച് /c/ , ഛ് /ch/ , ജ് /ɟ/ , ഝ് /ɟh/ , ഞ് /ɲ/ , ശ് /ʃ / , യ് /y/	

4.2 Data-driven approach

In data-driven approaches, visual features extracted from the mouth region of talking faces and viseme formed by clustering in the feature space. Both shape-based features and appearance-based features used as visual cues in the Malayalam language. Shape-based features use information from the speaker's lip contour. The geometric feature is the shape-based features used in this work. Appearance-based features deal with pixel information in the Region of Interest (ROI), thereby offer high computational complexity and is weak in capturing geometric variations when compared to shape-based feature. However, in a real-time application, appearance-based features show dominance over shape-based features which have the complexities related to accurate extraction of the lip contour. Discrete Cosine Transform (DCT) is the appearance-based features used in this work. Taking both methods together help in judging their reliability in the problem under study. K-mean clustering with the Gap statistic method is used to find the viseme set by clustering in the feature space.

4.2.1 Geometric visual features

Geometric features used in this study consist of outer lip width ($\mathrm{w}_{\mathrm{outer}}$), outer lip height ($\mathrm{h}_{\mathrm{outer}}$), inner lip width ($\mathrm{w}_{\mathrm{inner}}$), inner lip height ($\mathrm{h}_{\mathrm{inner}}$), outer lip area ($\mathrm{a}_{\mathrm{outer}}$), inner lip area ($\mathrm{a}_{\mathrm{inner}}$) and teeth area (t). These are extracted from the tracked lip contour as discussed in Sect. 2.

$$\mathbf{F}_{\mathbf{geometric}} = \left\{ \mathrm{w}_{\mathrm{outer}}, \mathrm{h}_{\mathrm{outer}}, \mathrm{w}_{\mathrm{inner}}, \mathrm{h}_{\mathrm{inner}}, \mathrm{a}_{\mathrm{outer}}, \mathrm{a}_{\mathrm{inner}}, \mathrm{t}_{\mathrm{area}} \right\}$$

The outer lip width and height are taken from the difference of x-coordinate of landmark points 1 and 11 and y-coordinate of landmark points 6 and 16, respectively as in Fig. 3a. Similarly, the inner lip width and height obtained from the difference of x-coordinate of cardinal points 21 and 29 and y-coordinate of cardinal points 25 and 33, respectively as in Fig. 3b. The outer lip area is the total number pixel points enclosed within the outer lip boundary as in Fig. 3c. The inner lip area is the oral cavity region, which measured by taking the total number of pixel points within the inner lip boundary as in Fig. 3d. The presence, absence, and area of teeth are direct indicators to distinguish many phonemes. The teeth area inside the convex hull of inner lip landmark points is computed after converting the pixels into the HSV colour space (Gritzman et al. 2015). Teeth pixels are segmented by applying a simple thresholding scheme to the pixels inside the inner lip as in Fig. 3e.

4.2.2 Discrete cosine transform (DCT) visual features

DCT is one of the old and still being used appearance-based visual feature extraction technique in literature. A two-dimensional DCT of an M-by-N image is represented as:

$$\mathrm{D}\,(\mathrm{i},\ \mathrm{j}) = \sum_{i=1}^{M}\sum_{j=1}^{N} I(i,j)\cos\left(\frac{(2i+1)\pi i}{2M}\right)\cos\left(\frac{(2j+1)\pi J}{2N}\right)$$

where I (i, j) is the grey-scale image of the ROI. The DCT return a 2-dimensional matrix having M*N coefficients. Most of the visually significant information and energy is concentrated in a few coefficients of DCT, which represent the low-frequency aspect of an image. Initially, the cropped speech frame of size 600 × 500 is further reduced to 64 × 64 for better implementation of the DCT algorithm. To avoid the curse of dimensionality, first, 20 coefficients per frames are selected from a 64 × 64 DCT coefficients in a zig–zag manner (McLaren and Lei 2015) starting from DC component [D (1, 1)].

4.2.3 Viseme set formation by clustering in the parametric space

Viseme set is formed by clustering in the feature vector space. Shape-based and appearance-based visual feature vectors are analyzed separately for 50 whole phonemes and 38 consonant phonemes. The geometric feature comprises of 7 numerical values per frame, and DCT features comprise of 20 numerical values per frame. Thus each phoneme is represented by a 35-dimensional geometric feature vector and 100-dimensional DCT feature vector respectively. The numerical representation of each phoneme is arranged horizontally in the feature

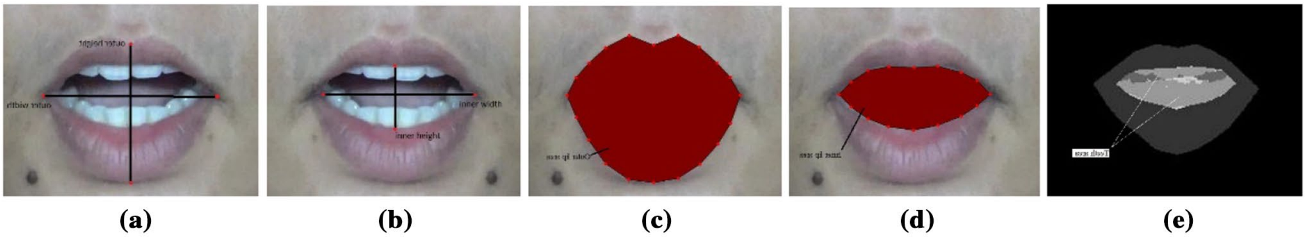


Fig. 3 Extraction of seven physical features from a frame

vector. An aggregate feature vector is created by horizontally concatenating the feature vector of 23 speakers. This feature vector was standardized for further analysis.

The final feature vector is fed into the Gap statistic method for determining optimum cluster number by making use of the k-mean algorithm for clustering purposes. K-mean algorithm is one of the simplest unsupervised learning algorithms which classifies the data set into a pre-defined number of clusters based on the centroid as in Jain (2010). The algorithm inputs are the dataset containing ‘n’ objects and pre-defined cluster number ‘k’. The algorithm of K-mean clustering given below:

1. The algorithms start with initial estimates for the K centroids, which can either be randomly generated or randomly selected from the data set.
2. Each centroid defines one of the clusters. In this step, each data point assigned to its nearest centroid, based on the squared Euclidean distance.
3. In this step, the centroids recomputed by taking the mean of all data points assigned to that centroid’s cluster.
4. The algorithm iterates between steps two and three until a stopping criterion met (i.e., no data points change clusters, the sum of the distances minimized, or some maximum number of iterations reached).

Due to the high correlation of mouth parameters for acoustically different phonemes, the clustering algorithm alone fails to estimate an optimum cluster value. From its birth to present, the Gap statistic method has revealed their strength in identifying optimum cluster number in an extremely correlated dataset. Gap statistic method compares the intra-class dispersion obtained from the given data with that of an appropriate reference distribution (Tibshirani et al. 2001). The mehodology of gap statistic work (using the natation from Tibshirani 2001) as follow,

Consider a datset $\{x_{ij}\}$ with $i = 1, 2, \ldots., n$ and $j = 1, 2, \ldots, p$, consists of p features measured on n independent observations, clustered into k clusters $C_1, C_2, \ldots, C_k$, where C_r denotes the indexes of samples in cluster r, and $n_r = | C_r|$. Let $d_{ii'}$ denotes the squared Euclidean distance between the observation i and i' ($d_{ii'} = \sum_j (x_{ij} - x_{i'j})^{2)}$. The sum of the pairwise distance D_r for all points in cluster r is:

$$D_r = \sum_{i,i' \in Cr} d_{ii'}$$

Let W_k be the within-cluster sum of squared distances from the clyster means as.

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$$

W_k decreases monotonically as the number of clusters k increases. For calculation the Gap function, Tibshirani et al. (2001) proposed to use the difference of the expected value of $\log(W*_k)$ of an appropriate null reference and the log (W_k) of the dataset,

$$Gap_n(k) = E*_n \log\left(W*_k\right) - \log(W_k)$$

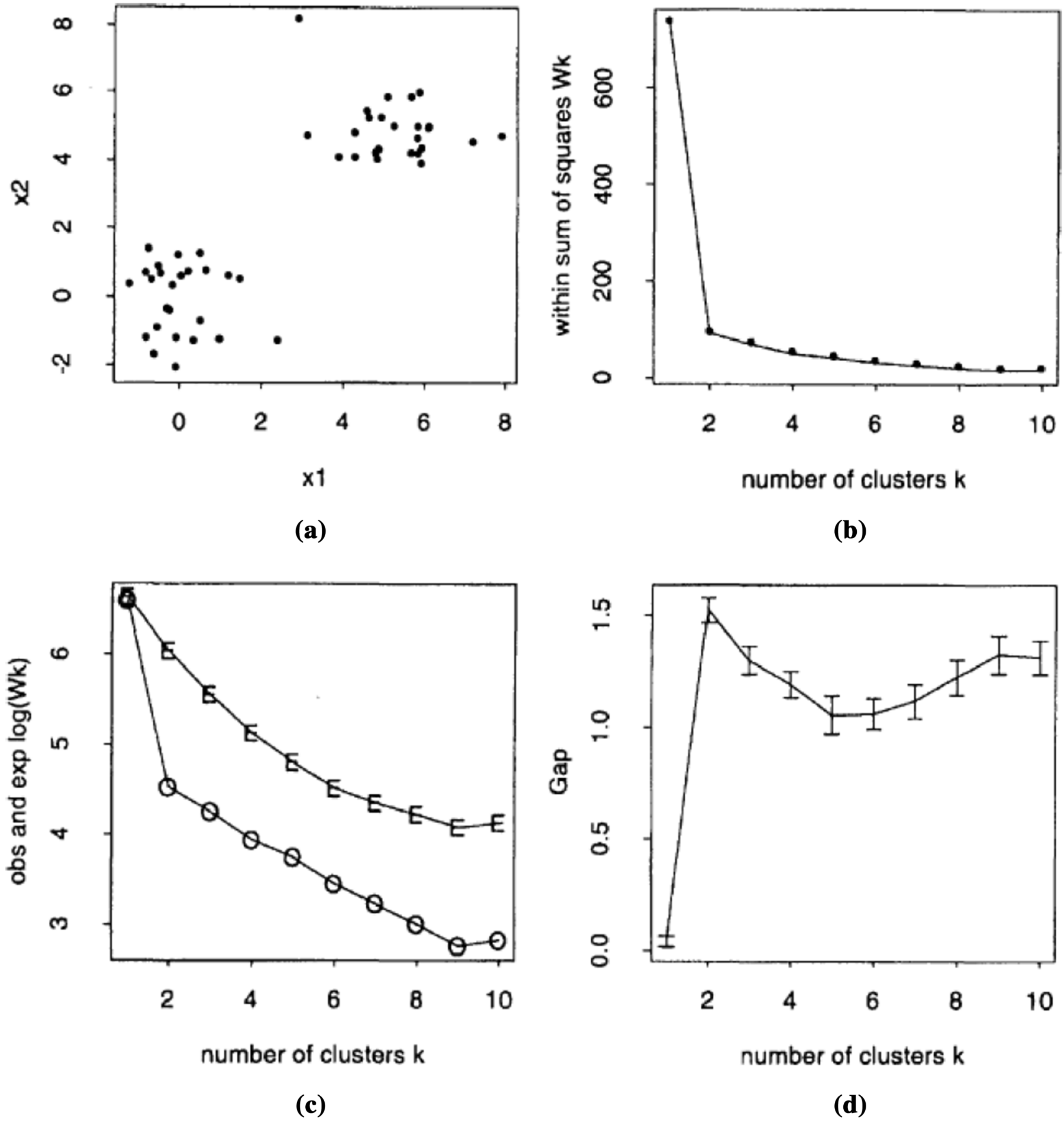
where $E*_n$ denotes the expectation of under a sample of size n from the reference distribution. Then proper number of clusters for the given data is the smallest k such that

$$Gap_n(k) \geq Gap_n(k+1) - s_{k+1}$$

Let s_k is the simulation error calculated from the standard deviation $sd\ (k)$ of B Monte Carlo replicates $\log(W*_k)$ according to the equation $s_k = \sqrt{1 + 1/B}\, sd(k)$ which is represented by vertical bar in Gap curve.

In short, A range of cluster groups is estimated using the k-means algorithm (or any other clustering algorithm), and the logarithm of within-cluster variance compared with the same measurement of an appropriate reference distribution of the data. The difference between these quantities (gap curve or gap function) provides the corresponding gap value for each cluster group, and these clusters show a fall at the point where the gap is maximum. Figure 4 reveals different Gap statistic steps of data using K-means clustering. For a well-separated dataset, the gap function exhibits a monotone behavior as in Fig. 4. However, for a highly correlated dataset, the gap function exhibits a non-monotone behavior, which directly educte us to inspect the whole gap curve instead of simply finding the cluster number with the maximum gap. The performance of the Gap statistic profoundly depends on the nature of the dataset/feature vector used. By considering the problem under study, features should select in such a way that it can display a high discriminating power between high correlated observables. In this work, the redundancy of selected features has been removed by considering a few feature coefficients that may deal with isuue under study. The optimum number of features needed to deal with the issue under study seriously remains an open research field.

Depending upon the property of dataset and underlying problem, it is better to study a reasonable range in the gap curve. Since analyzing the whole gap curve is troublesome and time consuming in estimating the optimum cluster number, especially for highly correlated data. For a highly correlated dataset and phoneme-to-viseme conversion problem, it is better to examine the gap curve between the clusters 10 to 20 Though the amount of viseme is language-dependent, the majority of the printed works have underlined this range in various languages. In this work too, 50 Malayalam phonemes are linguistically mapped to 14 viseme classes. Besides, for a straightforward interpretation, there will be a minimum of 2 phonemes can occupy in a single viseme



(a)

(b)

(c)

(d)

Fig. 4 A two cluster example: **a** data; **b** within sum of square function W_k; **c** functions log (W_k) (O) and $E^*_n\{\log(W_k)\}$ (E); **d** gap curve courtesy Tibshirani et al. 2001

class due to high correlation in the visual appearance of phonemes thereby creating a maximum of 25 viseme set. The same methodology is carried out in the rest of this paper. For clustering the feature vectors of 50 phonemes, gap curve analyzed, and the optimum cluster number is identified with maximum gap value in the cluster range 10 to 20, as shown in Fig. 5. For 50 Malayalam phonemes, based on geometric feature vector, the estimated the viseme set is 15 and shown in Table 6.

The diphthong vowel phonemes (ഔ-/au/), central vowel phonemes (അ-/a/, ആ-/a:/,), bilabial consonant phonemes (പ്-/P/,ബ്-/b/,ഭ്-/bh/,മ്-/m/), labiodental consonant phonemes (ഫ്-/ph/ ,വ്-/v/) and dental consonant phonemes (ത്-/t/, ഥ്-/th/, ദ്-/d/, ധ്-/dh/, ന്-/n̲/) velar consonant phonemes

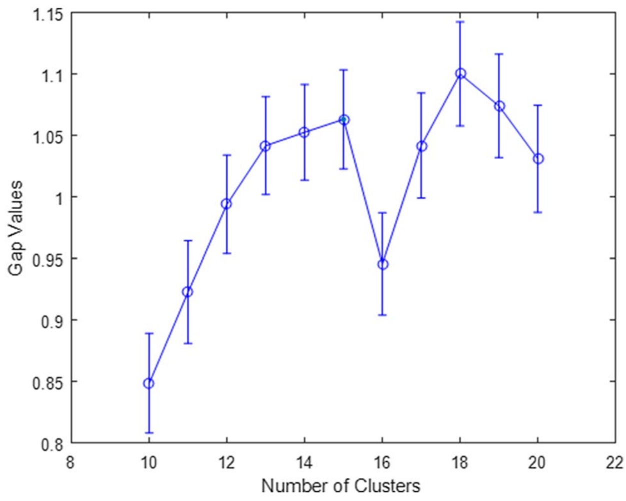


Fig. 5 Gap curve

(ക്-/k/, ഖ്-/kh/, ഗ്-/g/, ഘ്-/gh/, ങ്-/ŋ/) and most of the palatal consonant phonemes (ച്-/c/, ഛ്-/ch/, ജ്-/ɟ/, ഝ്-/ɟh/, ഞ്-/ɲ/) are grouped exactly in the same manner as in linguistic approach (Tables 4, 5). Due to lip contour similarity between front high vowel (ഇ-/i/, ഈ-/i:/) and front-mid vowel (എ-/e/, ഏ-/e:/) and back high vowel (ഉ-/u/, ഊ-/u:/) and back mid vowel (ഒ-/o/, ഓ-/o:/) they selected to the same viseme class rather than different classes as in linguistic approach. The remaining consonant phonemes distributed in such a way that it follows some traces of linguistic point of consonant phoneme cluster.

Phoneme-to-viseme mapping also studied for DCT feature vector with an estimated optimum cluster number equal to 16. The vowel phonemes are distributed precisely in the same fashion of linguistic approach but with overlapping of a diphthong phoneme (ഐ-/ai/) into the front-mid vowel phoneme class (എ-/e/, ഏ-/e:/). All most all consonant phonemes precisely as in the linguistic approach. The remaining consonant phonemes are randomly distributed just like in geometric feature map (Table 7).

After analyzing the data-driven approach, it is necessary to study consonant phoneme-to-viseme mapping since the consonant phoneme has failed to exhibit its distinguishing power except for some consonants. For grouping the consonant phonemes alone, the gap function is studied, with the geometric visual features, between the cluster ranges from 6 to 18 and the optimum cluster number estimated is 11 as shown in Table 8. In addition to bilabial and labiodental consonants, dental and palatal consonants have shown almost close resemblance with the linguistic mapping. However, the velar, glottal, alveolar, and retroflex consonants are shuffled within the class and between the classes, which make these consonant groups require more attention during visual speech recognition and synthesis applications.

Consonant phoneme-to-viseme mapping is also studied using DCT visual feature vector with an estimated optimum cluster number equal to 11. In this mapping, almost all viseme classes show a close resemblance with the linguistic mapping. While considering the complexities associated with problems, the DCT feature-based viseme mapping reveals a close correlation with the linguistic viseme than geometric feature-based viseme mapping (Table 9).

Table 6 Phoneme-to-viseme mapping based on the geometric feature vector

Viseme	Phonemes with IPA
1	റ്റ്-/r̠/
2	ഉ-/u/, ഊ-/u:/, ഒ-/o/, ഓ-/o:/
3	അ-/a/, ആ-/a:/, ഹ്-/h/
4	പ്-/p/, ബ്-/b/, ഭ്-/bh/, മ്-/m/
5	ല്-/l/, ള്-/ɭ/
6	ഫ്-/ph/
7	ത്-/t/, ഥ്-/th/, ദ്-/d/, ധ്-/dh/, ന്-/n̪/
8	യ്-/y/, ര്-/r/, ന്-/n/
9	ച്-/c/, ഛ്-/ch/, ജ്-/ɟ/, ഝ്-/ɟh/, ഞ്-/ɲ/
10	റ്-/ṛ/
11	ക്-/k/, ഖ്-/kh/, ഗ്-/g/, ഘ്-/gh/, ങ്-/ŋ/
12	വ്-/v/
13	ട്-/ʈ/, ഠ്-/ʈh/, ഡ്-/ɖ/, ഢ്-/ɖh/, ണ്-/ɳ/, ശ്-/ʃ /, ഷ്-/ʂ/, സ്-/s/, ഴ്-/ʐ/
14	ഔ-/au/
15	ഇ-/i/, ഈ-/i:/, എ-/e/, ഏ-/e:/, ഐ-/ai/

Table 7 Phoneme-to-viseme mapping based on the DCT feature vector

Viseme	Phonemes with IPA
1	ഒ-/o/, ഓ-/o:/
2	ക്-/k/, ഖ്-/kh/, ഗ്-/g/, ഘ്-/gh/, ങ്-/ŋ/
3	ഉ-/u/, ഊ-/u:/
4	പ്-/p/, ബ്-/b/, ഭ്-/bh/, മ്-/m/
5	അ-/a/, ആ-/a:/, ഹ്-/h/
6	ച് -/c/, ഛ്-/ch/, ജ്-/ɟ/, ഝ്-/ɟh/, ഞ്-/ɲ/
7	എ-/e/, ഏ-/e:/, ഐ-/ai/
8	ഇ-/i/, ഈ-/i:/
9	ട്-/ʈ/, ഠ്-/ʈh/, ഡ്-/ɖ/, ഢ്-/ɖh/, ണ്-/ɳ/,
10	ഺ്-/r̠/, ന്-/n/
11	ഔ-/au/
12	ശ്-/ʃ /, ഷ്-/ʂ/, സ്-/s/,
13	യ്-/y/, ര്-/r/, ല്-/l/,
14	ത്-/t/, ഥ്-/th/, ദ്-/d/, ധ്-/dh/, ന്-/n̪/
15	ള്-/ɭ/, ഴ്-/ʐ/, റ്-/ṛ/
16	ഫ്-/ph/, വ്-/v/

5 Allophone-to-viseme many-to-many mapping

While addressing the visual language, the phoneme-to-viseme many-to-one mapping must be needed to exhibit the high correlation bonding between phonemes and the visual look. However, this mapping admits viseme as a static visual address unit by eliminating the coarticulation effects of visual address. The presence of visual coarticulation effect creates a visibly different appearance for the same phoneme in a different context, which in turn creates a further subdivision in phoneme-to-viseme mapping. To put it differently, the visual appearance of a phoneme intensely depends not only on its articulation properties but also on the presence and nature of its neighbouring phoneme in the word or sentences. The contextual variation in Malayalam phonemes is modelled using allophonic characterization. For accurately describing the visual speech information in a different context, a comprehensive phoneme-to-viseme mapping needed, which ought to be an allophone-to-viseme many-to-many mapping. Only a few works were reported the coarticulation effects of visual speech by assembling the many-to-many mapping. Hilder et al. (2010) have described a novel method of segmenting the visual speech by capturing the patterns of behaviour of the articulators and clustered the behaviours that appear similar into a set of visemes, thereby obtaining a different viseme label for different allophones of a phoneme. Taylor et al. (2012) have modelled the coarticulation effects of visual speech by considering the motion of visual speech articulators rather than static mouth representation and animated a talking head. Mattheyses et al. (2013) have introduced a many-to-many phoneme-to-viseme mapping by using tree-based and k-means clustering approaches which ensures a more accurate description of visual speech when compared to phoneme based and many-to-one viseme based speech labels. Katsaggelos et al. (2015) discussed the challenges associated with audio-visual fusion, especially the need of audio-visual synchrony since range and directionality of coarticulation pattern differ across languages.

Linguistically there are 106 allophones which spread as 28 vowel allophones, three diphthong allophones and 75

Table 8 Consonant phoneme-to-viseme mapping based on the geometric feature vector

Viseme	Phonemes with IPA
1	പ്-/p/, ബ്-/b/, ഭ്-/bh/, മ്-/m/
2	ഴ്-/ʐ/, ന്-/n/
3	ഖ്-/kh/, ഗ്-/g/, ങ്-/ŋ/, ഹ്-/h/,
4	ഘ്-/gh/, ഡ്-/ɖ/, ന്-/n̪/
5	ക്-/k/
6	യ്-/y/, ര്-/r/, ല്-/l/, ള്-/ɭ/
7	ച്-/c/, ഛ്-/ch/, ജ്-/ɟ/, ഝ്-/ɟh/, ഞ്-/ɲ/, ഠ്-/ʈh/, ഢ്-/ɖh/, ണ്-/ɳ/
8	ഫ്-/ph/, വ്-/v/
9	റ്-/ṛ/, ഺ്-/r̠/
10	ട്-/ʈ/, ശ്-/ʃ /, ഷ്-/ʂ/, സ്-/s/
11	ത്-/t/, ഥ്-/th/, ദ്-/d/, ധ്-/dh/, ന്-/n̪/

Table 9 Consonant phoneme-to-viseme mapping based on the DCT feature vector

Viseme	Phonemes with IPA
1	ന്-/n/
2	ട്-/ʈ/, ഠ്-/ʈh/, ഡ്-/ɖ/, ഢ്-/ɖh/, ണ്-/ɳ/
3	ഫ്-/ph/, വ്-/v/ ഹ്-/h/,
4	പ്-/p/, ബ്-/b/, ഭ്-/bh/, മ്-/m/
5	ശ്-/ʃ /, ഷ്-/ʂ/, സ്-/s/, റ്റ്-/r̠/
6	ക്-/k/, ഖ്-/kh/, ഗ്-/g/, ഘ്-/gh/, ങ്-/ŋ/
7	ത്-/t/, ഥ്-/th/, ദ്-/d/, ധ്-/dh/, ന്-/n̺/
8	ച്-/c/, ഛ്-/ch/, ജ്-/ɟ/, ഝ്-/ɟh/, ഞ്-/ɲ/
9	റ്-/ṛ/, ഴ്-/ʐ/
10	യ്-/y/, ര്-/r/,
11	ല്-/l/, ള്-/ɭ/

consonant allophones as in Tables 1 and 2. Since the linguistic history of contextual variations of phonemes is not as researched from the Malayalam language, data-driven approach has to embrace for the building of many-to-many visual speech mapping. Geometric and DCT visual features are likewise captured from every allophone and has been clustered using K-mean algorithm and Gap statistic as in phoneme-to-viseme mapping. For clustering the geometric features of 106 allophones, gap curve is analyzed between the range 20 and 40, and the estimated cluster group is 32 and shown in Table 10.

The allophones of the very same phonemes are grouped randomly into 32 viseme groups. Some of the vowel allophones grouped in the exact same in addition to other vowel allophone classes as well as in consonant allophone groups. However, the most fascinating thing is distinct bilabial consonant allophones dispersed themselves without a crossover with additional anglophone bands, which shows its distinctive individuality in visual speech classification. Based on DCT visual features, 106 allophones grouped into 33 viseme groups with nearly similar clustering result as in allophone-to-viseme mapping based on geometric visual features. After assessing the data-driven approach, it is necessary to study vowel allophone-to-viseme mapping and consonant allophone-to-viseme mapping since both allophones radically connected as in Table 11.

For grouping the geometric features of 31 vowel allophones, the gap function is analyzed between the range 5 and 15, and the estimated cluster group is 11, as shown in Table 11.

The vowel allophones of the very same phonemes are grouped into different classes, thereby producing the tiniest many-to-many mapping. Only the single allophonic variation of ഔ-/au/ is showed a single clustering result when compared to the phoneme-to-viseme mapping, which directly highlights the sophistication of many-to-many mapping. The intricacy of consonant allophone-to-viseme mapping is studied by clustering the geometric visual features and assessing the gap curve over the range 20 to 35. Table 12 shows the clustering of 75 consonant allophones into 25 viseme groups.

Table 10 Allophone-to-viseme mapping based on the geometric feature vector

Viseme	Allophones with IPA	Viseme	Allophones with IPA
1	ദ്1-[ḍ], എ3-[ye]	17	മ്1-[m̪h], മ്4-[m̪h]
2	ക്6-[K], ഘ്1-[gh], ഹ്2-[H], ഖ്3-[Kh]	18	റ്റ്1-[d], ട്2-[ɽ]
3	ഉ3-[ɯ], ഔ1-[au]	19	അ2-[A], എ3-[E], ക്2-[kj], ഗ്1-[G], ങ്1-[ŋ
4	ത്4-[d̪], ഥ്1-[th]	20	ഢ്1-[ɖh]
5	ഗ്2-[g], റ്1-[ṛ], ങ്4-[ŋ>]	21	ദ്2-[ḍ], ന്2-[n]
6	മ്3-[m̪h], ഭ്1-[bh]	22	ഉ5-[ə*], ഘ്1-[gh], ത്3-[ð], ന്1-[n̺]
7	ല്1-[l], ഹ്1-[H], ള്1-[ɭ]	23	ട്3-[ʈ]
8	ണ്1-[ɳ]	24	ദ്1-[ḍ]
9	ര്1-[r]	25	ഖ്1-[kh], ങ്3-[ŋ<],], ങ്5-[ŋ']
10	ഉ1-[wu], ഉ2-[uw], ഉ7-[U], ഊ1-[wu:], ഊ2-[u]	26	ഇ1-[i], ച്4-[C], ഛ്1-[cʰ], ഛ്2-[Cʰ], ജ്2-[j], ഝ്1-[ɟh], ഞ്1-[ɲ] , ത്1-[t]
11	അ1-[ʌ], ആ2-[a], എ1-[ye], എ2-[ey], ഏ2-[er:], ഏ3-[e:], ഐ1-[ai], ഐ2-[ei], ക്4-[ɠ]	27	ഇ2-[yi], ങ്2-[ŋj], ഡ്1-[ɖ], ത്2-[t''], സ്1-[s], ഴ്1-[ʐ], റ്റ്2-[d], ന്1-[nh], ന്2-[nh]
12	ബ്1-[B]	28	ഒ1-[wO], ഒ2-[O], ഓ1-[wO:], ഓ2-[O]
13	പ്2-[p], ഫ്1-[ph], മ്2-[m̪h], വ്2-[v]	29	പ്1-[p], പ്3-[p], ബ്2-[B]
14	ക്1-[k], ഖ്2-[Kh], ച്1-[c], ധ്1-[dh], യ്1-[y]	30	ഇ3-[yi], ഈ1-[yi:], ഈ2-[i:], ക്3-[ɣ̥], ച്2-[ç̬], ച്3-[ɟ], ട്1-[ɖ], ഠ്1-[ʈʰ], ശ്1-[ʃ]
15	ആ1-[a:], ഏ1-[ye:]	31	വ്1-[w], പ്4-[P]
16	ഉ4-[ə], ക്6-[K], ട്4-[T], ഠ്2-[Tʰ]	32	ഉ6-[ɯv], ജ്1-[J], ഷ്1-[ʂ]

Table 11 Vowel Allophone-to-viseme mapping based on the geometric feature vector

Viseme	Allophone
1	ഇ1-[i], ഉ4-[wu]
2	ഏ2-[er:], ഏ3-[er:], ഐ1-[ai], ഐ2-[ai]
3	ഉ1-[wu], ഉ2-[wu], ഉ7-[wu], ഊ1-[wu:], ഊ2-[wu:], ഒ1-[wO], ഒ2-[wO], ഓ1-[wO:], ഓ2-[wO:]
4	ഉ5-[wu]
5	അ1-[ʌ], ആ1-[a], ആ2-[a], എ1-[ye], എ2-[ye]
6	അ2-[ʌ], ഇ2-[i], എ3-[ye]
7	ഔ1-[au]
8	ഏ1-[ye:]
9	ഉ3-[wu]
10	ഉ6-[wu]
11	ഇ3-[i], ഈ1-[yi:], ഈ2-[yi:]

Table 12 Consonant Allophone-to-viseme mapping based on geometric feature vector

Viseme	Allophones with IPA	Viseme	Allophones with IPA
1	ഷ്1-[ʂ]	14	പ്2-[p], മ്2-[m͟h]
2	ഗ്2-[G]	15	ക്1-[k], ച്3-[c], ജ്1-[ɟ], ട്3-[ɖ]
3	ഹ്2-[h], ഴ്1-[ʐ]	16	ബ്1-[B]
4	ക്3-[k], ഠ്2-[ʈʰ]	17	ഘ്1-[gh], ത്3-[t], ന്1-[n̪]
5	ക്1-[k], ച്1-[c], ഛ്1-[cʰ], ജ്2-[ɟ], ഝ്1-[ʄh], ഞ്1-[ɲ], ത്1-[t]	18	ഡ്1-[ɖ], ത്2-[t], സ്1-[s], ള്1-[ɭ], റ്റ്2-[t], ന്1-[nh], ന്2-[nh]
6	ത്4-[t]	19	ഫ്1-[ph], മ്3-[m͟h], വ്2-[w]
7	വ്1-[w], മ്1-[m͟h]	20	ഖ്1-[kh], ഢ്1-[ɖh], ദ്1-[ḍ]
8	ഹ്1-[H]	21	റ്റ്1-[d]
9	ഖ്2-[kh], ദ്2-[ḍ], ധ്1-[dh], ന്2-[n̪], യ്1-[y], റ്1-[ṛ]	22	പ്1-[p], പ്3-[p], പ്4-[p], ബ്2-[B], ഭ്1-[bh], മ്4-[m͟h]
10	ട്2-[ɖ]	23	ക്4-[k], ഖ്3-[kh], ട്4-[ɖ], ഠ്1-[ʈʰ]
11	ങ്2-[ŋ]	24	ശ്1-[ʃ]
12	ക്5-[k], ഗ്1-[G], ച്2-[c], ച്4-[c], ഛ്2-[cʰ], ഥ്1-[th], ര്1-[r]	25	ക്2-[k], ങ്1-[ŋ], ങ്3-[ŋ], ങ്5-[ŋ], ട്1-[ɖ], ണ്1-[ɳ]
13	ങ്4-[ŋ]	26	ല്1-[l]

The bilabial consonant allophones have revealed their particular distinguishing power nearly in exactly the exact same fashion as in the allophone-to-viseme mapping. The contextual variation of all other phonemes randomly distributed, which subsequently require prior focus in visual speech processing than compared to vowel allophone-to-viseme mapping.

6 Conclusion

In this paper, two facets of phoneme-to-viseme mapping assembled in the Malayalam language. For doing this work, an audio-visual Malayalam speech database made that consists of 23 trained speaker's uttering 50 isolated Malayalam phonemes along with 106 connected words comprising of all allophonic variations. In the first phase of the study, a phoneme-to-viseme many-to-one mapping is made according to the linguistic and data-driven approach. In the linguistic approach, 50 phonemes grouped into 14 viseme classes based on linguistic knowledge. In a data-driven approach, the viseme set is created by clustering the numerical representation of phonemes using the k-means algorithm and gap statistics. Geometric and DCT visual attributes used in the data-driven approach. Both features have classified the 50 phonemes into 14 viseme classes that are almost comparable with the linguistic classification of viseme as in Tables 4 and 5. This work also studied the consonant phoneme-to-viseme mapping based on geometric and DCT visual features which cluster 38 consonant phonemes into ten viseme classes.

To examine the coarticulation effect in visual speech, allophone-to-viseme many-to-many mapping generated. On account of the shortage of linguistic expertise, data-driven approach based allophone-to-viseme mapping created in the next phase of the work. The geometric and DCT visual features clustered the 106 allophones into 32 and 33 viseme classes respectively. Along with this, vowel alone and consonant alone, allophone-to-viseme mapping is also studied to acquire a comparative clustering plan of vowel and consonant allophones from the visual realm. In this work, unsupervised learning approach employed for producing the phoneme-to-viseme and allophone-to-viseme mapping. Developing these maps based on supervised learning algorithm will have a potential scope for the improvement in the present viseme set.

References

Aghaahmadi, M., Dehshibi, M. M., Bastanfard, A., & Fazlali, M. (2013). Clustering persian viseme using phoneme subspace for developing visual speech application. *Multimedia Tools and Applications, 65*(3), 521–541. https://doi.org/10.1007/s11042-012-1128-7.

Ahmad, N., Datta, S., Mulvaney, D., & Farooq, O. (2008). A comparison of visual features for audiovisual automatic speech recognition. *The Journal of the Acoustical Society of America, 123*(5), 3939. https://doi.org/10.1121/1.2936016.

Alexandre, D. S., & Tavares, J. M. R. S. (2010). Introduction of human perception in visualization. *International Journal of Imaging, 4*(10A), 60–70.

Alizadeh, S., Boostani, R., & Asadpour, V. (2008). Lip feature extraction and reduction for hmm-based visual speech recognition systems. In *International conference on signal processing proceedings, ICSP* (pp. 561–564). https://doi.org/10.1109/ICOSP.2008.4697195

Aschenberner, B., & Weiss, C. (2005). *Phoneme-viseme mapping for German video-realistic audio-visual-speech-synthesis* (pp. 1–11). Institut Für Kommunikationsforschung Und Phonetik, Universität Bonn.

Baswaraj, B. D., Govardhan, A., & Premchand, P. (2012). Active contours and image segmentation: The current state of the art. *Global Journal of Computer Science and Technology Graphics & Vision, 12*(11).

Bear, H. L., & Harvey, R. (2016). Decoding visemes: Improving machine lip-reading Helen L. Bear and Richard Harvey. In *Icassp 2016*, 2009–2013.

Bear, H. L., & Harvey, R. (2018). Comparing heterogeneous visual gestures for measuring the diversity of visual speech signals. *Computer Speech & Language, 52,* 165–190. https://doi.org/10.1016/j.csl.2018.05.001.

Bear, H. L., Harvey, R. W., & Lan, Y. (2017). *Finding phonemes: Improving machine lip-reading* (pp. 115–120). Retrieved from http://arxiv.org/abs/1710.01142

Binnie, C. A., Jackson, P. L., Montgomery, A. A. (1976). Visual intelligibility of consonants: A lipreading screening test with implications for aural rehabilitation. *Journal of Speech and Hearing Disorders, 41*(4), 530–539.

Biswas, A., Sahu, P. K., Bhowmick, A., & Chandra, M. (2015). VidTIMIT audio visual phoneme recognition using AAM visual features and human auditory motivated acoustic wavelet features. In *2015 IEEE 2nd international conference on recent trends in information systems, ReTIS 2015—Proceedings*, (2004) (pp. 428–433). https://doi.org/10.1109/ReTIS.2015.7232917

Blokland, A., & Anderson, A. H. (1998). Effect of low frame-rate video on intelligibility of speech. *Speech Communication, 26*(1–2), 97–103. https://doi.org/10.1016/S0167-6393(98)00053-3.

Bozkurt, E., Erdem, Ç. E., Erzin, E., Erdem, T., & Özkan, M. (2007). Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation. In *Proceedings of 3DTV-CON*. https://doi.org/10.1109/3DTV.2007.4379417

Brahme, A., & Bhadade, U. (2017). Phoneme visem mapping for Marathi language using linguistic approach. In *Proceedings—International conference on global trends in signal processing, information computing and communication, ICGTSPICC 2016* (pp. 152–157). https://doi.org/10.1109/ICGTSPICC.2016.7955288

Chitu, A. G., & Rothkrantz, L. J. M. (2009). Visual speech recognition automatic system for lip reading of Dutch. *Information Technologies and Control*, *year viii*(3), 2–9.

Damien, P., Wakim, N., & Egéa, M. (2009). Phoneme-viseme mapping for modern, classical arabic language. In *2009 international conference on advances in computational tools for engineering applications, ACTEA 2009* (Vol. 2(1), pp. 547–552). https://doi.org/10.1109/ACTEA.2009.5227875

Farooq, O., Datta, S., Shrotriya, M. C., Sarikaya, R., Pellom, B. L., John, H. L., et al. (2015). Er Er. *International Journal of Computer Applications, 1*(1), 1–4. https://doi.org/10.1109/ICASSP.2011.5947425.

Farooq, O., Upadhyaya, P., Farooq, O., Varshney, P., & Upadhyaya, A. (2013). *Enhancement of VSR using low dimension visual feature enhancement of VSR using low dimension visual feature*. (November). https://doi.org/10.1109/MSPCT.2013.6782090

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research, 11*(4), 796–804.

Franks, J. R., Kimble, J. (1972). The confusion of English consonant clusters in lipreading. *Journal of Speech and Hearing Research, 15*(3), 474–482.

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. https://doi.org/10.1007/s11760-014-0615-x.

Hazen, T. J., Saenko, K., La, C. H., & Glass, J. R. (2004). A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments. In *ICMI'04—Sixth international conference on multimodal interfaces* (pp. 235–242).

He, J., & Zhang, H. (2009). Research on visual speech feature extraction. In *Proceedings—2009 international conference on computer engineering and technology, ICCET 2009* (Vol. 2, pp. 499–502). https://doi.org/10.1109/ICCET.2009.63

Hilder, S., Theobald, B., & Harvey, R. (2010). In pursuit of visemes. In *Proceedings of the international conference on auditory-visual speech processing* (pp. 154–159). Retrieved from http://20.210-193-52.unknown.qala.com.sg/archive/avsp10/papers/av10_S8-2.pdf

Jachimski, D., Czyzewski, A., Ciszewski, T. (2018). A comparative study of English viseme recognition methods and algorithms. *Multimedia Tools and Applications, 77*(13), 16495–16532.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011.

Katsaggelos, A. K., Bahaadini, S., & Molina, R. (2015). Audiovisual fusion: Challenges and new approaches. *Proceedings of the IEEE, 103*(9), 1635–1653. https://doi.org/10.1109/JPROC.2015.2459017.

Lalitha, S. D., & Thyagharajan, K. K. (2016). A study on lip localization techniques used for lip reading from a video. *International Journal of Applied Engineering Research, 11*(1), 611–615.

Lander, J. (1999). *Read my lips: Facial animation techniques*.

Lee, S., & Yook, D. (2002). Audio-to-visual conversion using hidden Markov models. In *Lecture notes in computer science* (*Including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics*) (Vol. 2417, pp. 563–570).

Li, N., Lefebvre, N., & Lengellé, R. (2014, January). Kernel hierarchical agglomerative clustering: Comparison of different gap statistics to estimate the number of clusters. In *ICPRAM 2014—Proceedings of the 3rd international conference on pattern recognition applications and methods*, (pp. 255–262). https://doi.org/10.5220/0004828202550262

Lucey, P., & Potamianos, G. (2007). Lipreading using profile versus frontal views. In *2006 IEEE 8th workshop on multimedia signal processing*, *MMSP 2006* (pp. 24–28). https://doi.org/10.1109/MMSP.2006.285261

Madhulatha, T. S. (2012). An overview on clustering methods. *2*(4), 719–725. http://arxiv.org/abs/1205.1117

Mattheyses, W., Latacz, L., & Verhelst, W. (2013). Comprehensive many-to-many phoneme-to-viseme mapping and its application for concatenative visual speech synthesis. *Speech Communication, 55*(7–8), 857–876. https://doi.org/10.1016/j.specom.2013.02.005.

McLaren, M., & Lei, Y. (2015). Improved speaker recognition using DCT coefficients as features (pp. 4430–4434).

Meier, U., Stiefelhagen, R., Yang, J., & Waibel, A. (2000). Towards unrestricted lip reading. *International Journal of Pattern Recognition and Artificial Intelligence, 14*(5), 571–585. https://doi.org/10.1142/S0218001400000374.

Melenchón, J., Simó, J., Cobo, G., Martínez, E., La, A., & Llull, U. R. (2007). Objective viseme extraction and audiovisual uncertainty: Estimation limits between auditory and visual modes.

Miglani, S., & Garg, K. (2013). Factors affecting efficiency of K-means algorithm *2*, 85–87.

Mishra, A. N., Chandra, M., Biswas, A., & Sharan, S. N. (2013). Hindi phoneme-viseme recognition from continuous speech. *International Journal of Signal and Imaging Systems Engineering, 6*(3), 164–171. https://doi.org/10.1504/IJSISE.2013.054793.

Mohajer, M., Englmeier, K.-H., & Schmid, V. J. (2011). *A comparison of Gap statistic definitions with and without logarithm function*. Retrieved from http://arxiv.org/abs/1103.4767

Montgomery, A. A., & Jackson, P. L. (1983). Physical characteristics of the lips underlying vowel lipreading performance. *Journal of the Acoustical Society of America, 73*(6), 2134–2144. https://doi.org/10.1121/1.389537.

Morade, S. S. (2016). Visual lip reading using 3D-DCT and 3D-DWT and LSDA. *International Journal of Computer Applications, 136*(4), 7–15.

Morade, S. S., & Patnaik, S. (2014). Lip reading by using 3-D discrete wavelet transform with Dmey wavelet. *International Journal of Image Processing, 8,* 384–396.

Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., et al. (2000). *Audio visual speech recognition* (No. REP_WORK). IDIAP.

Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., & Ogata, T. (2015). Audio-visual speech recognition using deep learning. *Applied Intelligence, 42*(4), 722–737. https://doi.org/10.1007/s10489-014-0629-7.

Puviarasan, N., & Palanivel, S. (2011). Lip reading of hearing impaired persons using HMM. *Expert Systems with Applications, 38*(4), 4477–4481. https://doi.org/10.1016/j.eswa.2010.09.119.

Rajavel, R., & Sathidevi, P. S. (2009). Static and dynamic features for improved HMM based visual speech recognition. In *Proceedings of the first international conference on intelligent human computer interaction* (pp. 184–194). https://doi.org/10.1007/978-81-8489-203-1_17

Saitoh, T., & Konishi, R. (2010). A study of influence of word lip-reading by change of frame rate. *Word Journal of the International Linguistic Association* (pp. 400–407).

Sarma, M., & Sarma, K. K. (2015, May). *Recent trends in intelligent and emerging systems* (pp. 173–187). https://doi.org/10.1007/978-81-322-2407-5

Seko, T., Ukai, N., Tamura, S., & Hayamizu, S. (2013). Improvement of lipreading performance using discriminative feature and speaker adaptation. In *Avsp*.

Setyati, E., Sumpeno, S., Purnomo, M. H., Mikami, K., Kakimoto, M., & Kondo, K. (2015). Phoneme-viseme mapping for Indonesian language based on blend shape animation. *IAENG International Journal of Computer Science, 42*(3), 1–12.

Stewart, D., Seymour, R., & Ming, J. (2008). Comparison of image transform-based features for visual speech recognition in clean and corrupted videos. *Eurasip Journal on Image and Video Processing, 2008*(2008), 1–9. https://doi.org/10.1155/2008/810362.

Sui, C., Bennamoun, M., & Togneri, R. (2016). Visual speech feature representations: recent advances. In *Advances in Face Detection and Facial Image Analysis* (pp. 377–396). Cham: Springer.

Taylor, S. L., Mahler, M., Theobald, B. J., & Matthews, I. (2012). Dynamic units of visual speech. In *Computer animation 2012—ACM SIGGRAPH/eurographics symposium proceedings*, *SCA 2012*, (pp. 275–284).

Taylor, S., Theobald, B. J., & Matthews, I. (2015). A mouth full of words: Visually consistent acoustic redubbing. In *ICASSP*, *IEEE international conference on acoustics, speech and signal processing—proceedings*, *2015–August* (pp. 4904–4908). https://doi.org/10.1109/ICASSP.2015.7178903

Tibshirani, R., Walther, G., Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*(2):411–423.

Upadhyaya, P., Farooq, O., Abidi, M. R., & Varshney, P. (2015). Comparative study of visual feature for bimodal hindi speech recognition. *Archives of Acoustics, 40*(4), 609–619. https://doi.org/10.1515/aoa-2015-0061.

Varshney, P., Farooq, O., & Upadhyaya, P. (2014). Hindi viseme recognition using subspace DCT features. *International Journal of Applied Pattern Recognition, 1*(3), 257. https://doi.org/10.1504/ijapr.2014.065768.

Websdale, D., & Milner, B. (2015). Analysing the importance of different visual feature coefficients. *Faavsp, 3,* 137–142.

Xiaopeng, H., Hongxun, Y., Yuqi, W., & Rong, C. (2006). A PCA based visual DCT feature extraction method for lip-reading. In *Proceedings—2006 international conference on intelligent information hiding and multimedia signal processing*, *IIH-MSP 2006* (December 2006) (pp. 321–324). https://doi.org/10.1109/IIH-MSP.2006.265008

Yu, D., Ghita, O., Sutherland, A., & Whelan, P. F. (2010). A novel visual speech representation and HMM classification for visual speech recognition. *IPSJ Transactions on Computer Vision and Applications, 2,* 25–38. https://doi.org/10.2197/ipsjtcva.2.25.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.