A Novel Approach for Allocating Mathematical Expressions to Visual Speech Signals

Mohammad Hossein Sadaghiani¹, Niusha Shafiabady¹, and Dino Isa¹

Abstract
In this article, visual speech information modeling analysis by explicit mathematical expressions coupled with words’ phonemic structure is presented. The visual information is obtained from deformation of lips’ dimensions during articulation of a set of words that is called visual speech sample set. The continuous interpretation of the lips’ movement has been provided using Barycentric Lagrange Interpolation producing a unique mathematical expression named visual speech signal. Hierarchical analysis of the phoneme sequences has been applied for words’ categorization to organize the database properly. The visual samples were extracted from three visual feature points chosen on the lips via an experiment in which two individuals pronounced the aforementioned words. The simulation results show that each individual word can be represented by a mathematical expression or visual speech signal whereas the sample sets can also be derived from the same mathematical expression, and this is a significant improvement over the popular statistical methods.

Keywords
viseme, visual speech signal, Barycentric Lagrange Interpolation

Introduction
The audiovisual speech recognition and visual speech synthesizers are two interfaces for human and machine interaction (Chin, Seng, & Ang, 2012). Such interaction relies on the analysis and synthesis of both audio and visual information, which humans use for face-to-face communication. It has been shown that the visual cues of speech also enhance the transparency of speech when they are degraded by environmental noise (Sumby & Pollack, 1954).

The visual speech information is the bimodal component of audio speech signal. These modalities are strongly correlated that could affect the perception of lip movement and audio speech signal (McGurk & MacDonald, 1976). In audiovisual speech recognition systems, the visual information is used as complementary tool for enhancing the perception of speech signal modality or independently for lip reading. The visual recognition module is called interchangeably as lip-reading, visual speech recognition, speech reading, or visual-only automatic speech recognition (Petajan, 1984; Potamianos, Neti, Luettin, & Matthews, 2004).

In lip-reading systems, the dynamics of visual speech in image sequences are extracted by focusing on the appearance of the articulator organs like lips’ geometry. Further processing could be conducted to be compatible for fusing to the audio-only automatic speech recognition modality.

A basic component related to phoneme in the visual domain was determined and called viseme (visual phoneme; Fisher, 1968). Since then, no globally accepted model has been suggested for formulating the visual speech components. The concept of visual phoneme does not suggest an explicit definition of lips’ structure during phoneme utterance. The visemes are formed based on human perceptions which are categorized using confusion matrix where the most accurately detected visemes form a phoneme–viseme table (Williams, Rutledge, Garstecki, & Katsaggelos, 1997). The deficiency of this method can be observed by the fact that there are various phoneme–viseme tables used (Goldschen, Garcia, & Petajan, 1994; Hazen, Saenko, La, & Glass, 2004; Jiang, Alwan, Auer, & Bernstein, 2001). The next drawback of the viseme concept was the lack of accuracy in visual phonemes perception caused by human interference. In other words, depending on the observers, whether they are hearing impaired or not, the results were changing. In addition, the concept of viseme represents the visual information

¹University of Nottingham, Malaysia Campus, Semenyih, Malaysia

Corresponding Author:
Niusha Shafiabady, School of Electrical and Electronic Engineering, University of Nottingham, Malaysia Campus, Semenyih 43500, Malaysia.
Email: niusha.shafiabady@nottingham.edu.my
as discrete components. Therefore, the continuity of visual speech components must be conducted by complementary methods (Cohen & Massaro, 1993).

To eliminate the additional methods for preserving the continuity of visemes, the most dominant approaches are statistical and probabilistic models (e.g., Hidden Markov Model [HMM]; Heracleous, Tran, Nagai, & Shikano, 2010; Kosmopoulos & Chatzis, 2010; Potamianos & Graf, 1998; Saenko, Livescu, Glass, & Darrell, 2009). The advantage of this approach is its ability to apply modifications and adjustments to the model using optimization methods and training algorithms (Huang & Povey, 2005; Lee & Park, 2006). However, those approaches do not have the ability to formulate the relations between viseme sequences. Therefore, currently there is no standard method which is globally accepted for representing the lips’ movement during articulation.

In Reyet and Benoît (1998), a 3D lip model was fitted to lip images, and the amount of similarities between the feature points on the model and actual lips were measured. Their results did not conclude the dynamic of articulating lips with a mathematical expression. In the suggested lip-reading system by Potamianos and Graf (1998), the visual features related to the width, height, and area of lip were tracked over frame or time domain during articulating a sequence of digits. Such trajectories of visual features were used for training a Hidden Markov Model that was not mathematically formulated. The acquired data (from movement of the lips) in Lucero (2002) was formulated by a sixth-order B-spline curve (Mortenson, 1997). The main disadvantage of using a B-spline curve for formulating the lips’ movement is the lack of definition of a single expression for the sample sets as B-spline curve is defined on segmented intervals (Mortenson, 1997). The visual information was extracted from tongue tip, tongue blade, tongue dorsum, velum, upper and lower lip, and lower jaw by Electromagnetic Articulography (Ananthakrishnan & Engwall, 2008).

The method suggested in Birkholz, Kröger, and Neuschäfer-Rube (2011) is a time-variant dynamic system used for modeling the articulatory organ to relate facial parts mentioned above by using Electromagnetic Articulography approach. A combination of phoneme was chosen with normal and low speaking rates. The input–output relations of the systems were obtained by solving differential equation in time domain. Although the proposed method tried to allocate a transfer function to describe the dynamics of lips and the related organs, finding the solution of the system in time via differential equations caused computational overhead and do not represent a single expression for the lip movement that is considered as a disadvantage for this method.

The extraction of feature points using the appearance-based approach in Zhao, Barnard, and Pietikainen (2009) was suggested using a statistical visual feature modeling by concatenating the image sequence portions in block volume. Applying an image processing approach called Local Binary Pattern (LBP), the lips’ textures between successive images in gray scale are used and transformed to binary codes. Afterward, the LBP is applied on the other two orthogonal planes to the image plane, which codes the spatial and temporal axis. The histogram of these binary orthogonal planes was calculated and represented as the visual feature. By concatenating of the histograms, two important characteristics as lips’ appearance and also the motion of lips are extracted. The sequences of histograms were joined to represent the overall statistical property of lip movement. In other words, their method not only extracts the visual feature but also expresses the deformation of visual features dynamically. Interpreting the dynamic of lips is determined based on analyzing the extracted data as appearance-based methods including statistical models (Adjoudani & Benoît, 1996; Bregler & Konig, 1994; Erber, Sachs, & DeFilippo, 1979) and shape-based methods or geometric approaches (Chiou & Hwang, 1997; Rogozan & Delèglise, 1998; Teissier, Robert-Ribes, & Schwartz, 1999) and a combination of appearance-based and shape-based methods (Cootes, Edwards, & Taylor, 1998). The necessity for suggesting a globally defined visual phoneme concept supports the mathematical modeling of the visual information. Therefore, the focus of this article is on extracting the geometry of lips during articulation to suggest a mathematical model of lips’ dynamics as there is no specific attempt for expressing the visual speech data during articulation by an explicit mathematical formula. In this article, a straightforward method for formulating the visual speech data is represented, and the above-mentioned data is extracted from transcribed words which are designed based on phonemic structures.

There are two varieties of human–machine interactions, where in the first one, the machine plays a role as interface in human-to-human interaction (audiovisual speech recognition) or animation (visual speech synthesis). In the audiovisual automatic speech recognition (AVASR) systems, the visual information is the input of identification for enhancing the perception of speech signal modality. This visual recognition module is called interchangeably as lip-reading, visual speech recognition (Petajan, 1984), speech reading, or visual-only automatic speech recognition (Potamianos et al., 2004). In the lip-reading systems, by focusing on the appearance of the lip’s geometry or pixels in image sequences, the dynamic of articulating lip is extracted. This information has to be processed to be compatible for fusing to the audio-only automatic speech recognition (ASR) modality. The usage of viseme appears in this stage. However, in animation or visual speech synthesis, the speech signal (Massaro, Beskow, Cohen, Fry, & Rodriguez, 1999) or transcribed information of words (Ezzat & Poggio, 1998) is used for lip’s movement identification. In both cases, the lip movement is interpreted from phoneme–viseme table. Furthermore, in audiovisual speech processing, the visual expressions can also be used as audio speech anticipators. The anticipation effect has been represented in Kim and Davis (2003), where in Conrey and Pisoni (2003), the asynchrony of visual speech
and audio speech information perception are examined. It has been shown that the tolerance of perceivers to visual speech cues preceding the audio speech cues is higher.

Referring to viseme, its framework has not been explained systematically. The main issue in viseme concept was lack of globally accepted definition for relation between phoneme and its visual appearance, for example, there are varieties of phoneme–viseme tables. The next drawback of the viseme concept is the lack of accuracy in perception of visual phonemes due to dependency on human perception of visual phonemes. In other words, depending on the observers whether they are hearing impaired or not, the results will vary. The next important issue is lack of explicit definition of the lip’s deformation to make a standard model of articulating lips. The concept of visual phoneme does not suggest an explicit definition of lip’s structure during phoneme utterance. In addition, the main factor of human perception in evaluating and selecting viseme could vary from one observer to another. Therefore, there is not any globally agreed and standardized phoneme–viseme table reference.

To interpret the dynamic of lips, many approaches have been employed. These approaches belong to three main groups depending on analyzing the extracted data as appearance-based methods including statistical methods (Adjoudani & Benoît, 1996; Bregler & Konig, 1994; Erber et al., 1979) and shape-based methods or geometric approaches (Chiou & Hwang, 1997; Rogozan & Deléglise, 1998; Teissier et al., 1999) and combinations of appearance-based and shape-based methods (Cootes et al., 1998). The necessity for suggesting a globally defined visual phoneme concept suggests expressing of the visual information by mathematical models. Therefore, the focus of this article is on extracting the geometry of lip during articulation to suggest a mathematical model for lips’ dynamics.

A 3D lip model was fitted to lips’ image (Revéret & Benoît, 1998), and the amount of similarities between the feature point on the model and actual lip has been measured. Their results did not conclude the dynamic of articulating lip with a mathematical expression. In the suggested lip-reading system by Potamianos and Graf (1998), the visual features related to the width, height, and area of lip were tracked over frame/time domain during articulating a sequence of digits “81926.” Such trajectories of visual features were used for training a HMM model but were not mathematically formulated. The acquired data (from movement of lip) in Lucero (2002) was formulated by B-spline cure (Mortenson, 1997). The main disadvantage of using B-spline for formulating the lip’s movement was lack of defining a single expressions for the sample sets as B-spline curve functions on segmented intervals. The extraction of feature points using the appearance-based approach in Zhao et al. (2009) was suggested by a statistical visual feature modeling by concatenating the image sequence portions in block depicted volumes. Using an image processing approach called Local Binary Pattern (LBP), the lip texture between successive images in gray scale are used to transform to binary codes. Afterward, the LBP is applied to the other two orthogonal planes to the image plane, which codes the spatial and temporal axis. The histogram of these binary orthogonal planes was calculated and represented as visual feature. By concatenating of the histograms, the two important characteristics (lips’ appearance and also the motion of lip) are extracted. The sequences of histograms are joined to represent the overall statistical property of lip’s movements. In other words, their method not only extracts the visual feature but also expresses the deformation of visual features dynamically. The method suggested in Birkholz et al. (2011) is a time-variant dynamic system used for modeling the articulatory relation of facial parts as tongue tip, tongue blade tongue, jaw, dorsum, upper lips, and lower lips by using electromagnetic articulography approach. The combination of phoneme was chosen as [CVCVCVCV] sequence with normal and low speaking rates. The input–output relation of the systems was obtained by solving differential equations in time domain. Although the proposed method tried to allocate a transfer function to describe the dynamics of lips and the related organs, the solution of the system in time via differential equations was computationally expensive and does not represent a single expression for the lip movement. According to the literature, there is not any specific attempt for expressing the visual speech data during articulation by an explicit mathematical formula. In this article, a straightforward method for formulating the visual speech data is represented. The visual speech data is extracted from a transcribed words which are designed based on phonemic structure.

Having seen the need for a standard approach to model the visual phoneme, the emphasis of this research is on introducing a mathematical model for human visual speech. In this article, a novel method has been introduced for the analysis of the visual speech that has been developed based on both lips’ movements and linguistic clues in English language. The visual speech information was related to a set of words chosen based on phoneme sequences from a transcribed database.

This article is organized as follows. “Visual Feature Determination” section explains visual feature determination procedure. “Corpus Design” section discusses corpus design. Visual speech sample extraction is described in “Visual Speech Sample Extraction” section. The modeling of visual speech signals is shown in “Visual Speech Signal” section, and the results are given in “Results and Discussion” section. Finally the conclusion is discussed in “Conclusion” section.

**Visual Feature Determination**

In this section, a visual sample set which describes the lip geometry is defined. The process is also referred as parameterizing the lip. The visual samples extracted from these visual feature points are the upper lip position, the lower lip position, and the corner lip position. These positions are
measure the visual features according to the Cartesian coordinate system that is defined within the lips’ geometry. The displacement of the visual features in Equations 1 to 3 leads to defining the visual speech sample sets as follows.

\[ FP_u(i) = \{ y_{0u}, y_{1u}, y_{2u}, \ldots, y_{wu} \}, \]

\[ FP_l(i) = \{ y_{0l}, y_{1l}, y_{2l}, \ldots, y_{wl} \}, \]

\[ FP_c(i) = \{ y_{0c}, y_{1c}, y_{2c}, \ldots, y_{wc} \}, \]

where \( w \) is representing a word with specific number of frames \( m \). These sample sets are used for describing lips’ movements during articulating a set of words. Now that the lips are parameterized, we move on to the next section where a method for designing the transcribed data is represented.

**Corpus Design**

The arrangements of phonemes in transcribed form (corpus) are also considered as one of the important issues in visual speech analyzers and synthesizers. It addresses the structures of phonemes used in lip-reading systems and visual animation of speech. There are several different corpora that differ in the structure of phonemes for generating lip movements. The structures of the corpora are shown in Kim and Davis (2003) where \( C \) and \( V \) refer to consonant and vowel phonemes, respectively. Each set corresponds to specific combinations of isolated phonemes.

In the first category of transcribed data, the isolated phoneme was used for lip reading (Petajan, 1984) or animation (Bregler, Covell, & Slaney, 1997). In the second scenario, a specific arrangement of vowel or consonant was fixed for visual analysis of speech. A corpora developed by Montgomery (1980) used the /CVVC\( C_2\)/ synthesizer while Adjoudani and Benoit (1996) used /\( VC_1 CV_1\)/ corpus. The random combination of phonemes as /aCa/ was addressed in the corpora structure by Su and Silsbee (1996). The phoneme arrangements as a part of a French sentence “C’est pas / VCVC\( V_2\)/” was also examined by Revéret and Benoit (1998). Czup (2000) used triphone sequences /\( VC_1 CV_2\)/ and /\( VC_2 CV_1\)/ corpora. For a Brazilian Portuguese facial animation system proposed by De Martino, Magalhaes, and Violaro (2006), the corpora consisted of two types of phonemes combinations as /CV_1 CV_2/ and /\( VC_1 V_1\)/ while the /VCV/ structure was used by Cox, Harvey, and Newman (2008) and Alotmany, Boston, Li, Shaiman, and Durrant (2010). This model was used for training a HMM in ASR system with different four conditions under seven levels of noise.

The suggested structure of corpus of Castilian Spanish used by Melenchón, Martínez, De La Torre, and Montero (2009) consisted of /CVVC\( V_2\)/. Their purpose for using such structure was supported by a strong statement, which claims...
Sadaghiani et al.

Table 1. The Structure of Phonemes Used as Corpora.

| Corpora                | Authors                          | Language          |
|------------------------|----------------------------------|-------------------|
| /CV CV/                | Montgomery (1980)                | English           |
| /V CV_1 CV_2/         | Adjoudani and Benoît (1996)      | English           |
| /aCa/                 | Su and Silsbee (1996)            | English           |
| “C’est pas /VCVCVz/?” | Revéret and Benoît (1998)        | French            |
| /V CV_1 / and /C CV_1/| Czup (2000)                      | Hungarian         |
| /CV CV_1 / and /V CV_1/| De Martino, Magalhaes, and Violaro (2006) | Brazilian Portuguese |
| /VCV/                 | Cox, Harvey, and Newman (2008)   | English           |
| /CVCV/                | Melenchón, Martínez, De La Torre, and Montero (2009) | Castilian Spanish |
| /VCV/                 | Alothmany, Boston, Li, Shaiman, and Durrant (2010) | English |
| /CVCVCVCV/            | Birkholz, Körger, and Neuschaefer-Rube (2011) | Not stated |

Figure 2. The designing steps toward extracting a limited set of words. Note. TIMIT = Texas Instrument and Massachusetts Institute of Technology.

more than 80% of Castilian Spanish words flow /CV/ structure (de Vega, Álvarez, & Carreiras, 1992). A suggested phoneme arrangements by Birkholz et al. (2011) produced /CVCVCVCV/ sequences. These methods are also noted in Table 1 where accordingly no evidence for choosing such arrangements of phoneme sequences was presented. The main disadvantage of such corpora arrangement is ignoring the actual sequences of consonants and vowels in existing words, in contrary with the above considered sequences.

Hierarchical analysis of the phoneme sequences has been applied for words’ categorization to organize the database properly that covers the above-mentioned disadvantage. The sequences of phonemic symbols are the subject of such analysis. Driving visual speech signals for visual speech samples will be more comprehensive if the signals are produced from a specific family of words. Accordingly, the most frequent phonemic alphabets that lead to the most frequent sets of words which have similar phoneme sequences can be extracted. The phonemic structures make the words families. This procedure is called phoneme-based analysis of words in the current study.

In reality, words have specific uniquely defined sequences of phonemes. Generating the combination of phoneme sequence starts from the beginning of a word. This eliminates the need to study the random combination of phonemes. Obviously, many words share the same phoneme sequence in their initial parts. One of the significant advantages of phoneme-based analysis of words is the ability of controlling the mathematical expressions in the visual domain. In this work, the common phonemes in the initial parts of the words are found based on the search in three sequences of phonemes (triphones), and this is the basis of designing the corpus.

In Figure 2, detail about the hierarchical approach for obtaining a limited set of words is illustrated. One of the most well-known sources of ASR, which provides varieties of the audio-only databases bodies, is the Linguistic Data Consortium (LDC). Among the databases, the Texas Instrument and Massachusetts Institute of Technology (TIMIT) Acoustic-Phonetic Continues Speech Corpus is chosen for this study. This database was widely used in the ASR systems like Hidden Markov Model Toolkit (HTK).

The hierarchical analysis is applied to entire corpus to derive more information about the behavior of phonemes in the corpus.

This concept provides the relation between the mathematically expressed visual speech signals called visual words and the meaningful combinations of sounds that are marked by words. Consequently, the mathematical expressions, which
would be driven for the visual words, are also categorized as a group. Therefore, the words are grouped mathematically. In mathematical form, the difference between each member of group will be revealed after the third phoneme. Therefore, the visual speech signals demonstrate a systematic relation between a member of the word’s family and its family members. This method could be used in lip-reading and also animation systems. Toward establishing the signals expression for structural words, a set of words were chosen by the criteria of maximum phonemic sequence occurring among all the transcribed words. This method can be applied to any word in dictionary, but for this study, a collection of words is chosen in English language for synthesizing the visual word database.

Adaptation Phase

In the field of speech processing, it is possible to select the appropriate viseme table that has been chosen as the visual domain representation reference. Selecting appropriate phoneme–viseme mapping table for analysis of phonemic structure would be beneficial in deriving the practical visual speech signal expressions that are adaptable to audiovisual speech recognition systems. In other words, the method of deriving visual speech signals could be applied to the speech recognizer system as its collaborating visual domain. According to Potamianos et al. (2004), a phoneme set used in HTK was selected as the phoneme–viseme mapping table. In this mapping scheme, the many-to-one strategy was applied based on the manner of articulation. The consonants consist of eight subgroups of phonemes that have visual similarities during articulation, while vowels are categorized into four subgroups.

Phoneme-Based Hierarchical Analysis

In this part, the hierarchical analysis is used for studying the combination of phonemic components in the words. In the hierarchical analysis of words phonemic structure, the goal is grouping all possible sequences of consonants and vowels. This process schematically is depicted in Figure 3.

Words Phonemic Pattern Analysis

The hierarchical analysis provides the information for systematic analysis of organized data. Using this approach to the phonemic representation of words, the hierarchical probability information of phonemes sequences is extracted as they appear in words. A decision tree is defined in sequentially arranged levels. The nodes relations were defined by connecting lines from their current level to the other nodes in the successive level. The nodes represent a particular state, group, event, or position in each level that exists in the data set. The connecting lines represent conditional rules that define the transition among the nodes between two consecutive levels. These rules are stated by the maximum numbers of repetitions mapped to phonemes in each level. In this work, the nodes of the decision tree are denoted by the consonants and vowels. The tree representation of phonemes sequences is shown in Table 2 where 13 words are selected and categorized having the same phonemic sequence for three levels.

Visual Speech Sample Extraction

The practical phase of this work was represented in Figure 4. After designing a corpus, the test subjects were asked to articulate the words. A camera captured their lips’ movements. The lips’ geometry was parameterized on the region of interest by three static feature points located on upper, lower, and corner outer contour of lip. Therefore, each word has three sets of samples in visual domain. The movement of these feature points can describe the lips’ movement. The visual data was extracted automatically and manually.

The speakers were two non-native women aged between 25 and 31 with no speech disorders or hearing disabilities. The visual speech sample sets $FWm(i), FPWm(i),$ and $FPWm(i)$ were extracted manually from the upper $fp^u(i)$, lower $fp^l(i)$, and the corner $fp^c(i)$ visual feature points on two speakers’ lips demonstrated in Figure 5a to 5c, respectively, for $i = \{0,1,2,\ldots,F_{Wm} - 1\}$ where $F$ is substituted by $F_{Wm}$.

Visual Speech Signal

The Lagrange (Lagrange-Newton) interpolation was a method of polynomial formulation, which constructs a continuous-time polynomial $L(x)$ from $N$ number of samples taken from a function $y = f(x)$ and $i \in \mathbb{N}$ over an equidistance interval $x$, where $x_i \in [0 : N - 1]$. The corresponding amplitudes of samples are defined as $a_i = f_i$. The variable $x$ was defined via changing the polynomial’s sampling frequency $\beta$. More specifically, by recalling the definition of continuous-time interval (where $\beta \rightarrow \infty$), the interval $x$ is defined as

$$x = \left\{x_i\right\}, x_i = \frac{i}{\beta}, i = 0, 1, 2, \ldots, \beta(N - 1) \in \mathbb{N} \quad (7)$$

Figure 3. Searching for word’s phonemic rules.
Table 2. The Tree Representation of Phonemes Sequences Categorizing All 13 Words as Branches of One Root.

| Level | Word groups |
|-------|-------------|
| 1     | /s/ /ih/ /m/ P /l/ /ax/ /t/ — — — — — — Simple |
| 2     | — — — — — — — Simpler |
| 3     | /iy/ — — — — — — — Simplest |
| 4     | /el/ — — — — — — — — Simply |
| 5     | /ax/ /th/ /eh/ /t/ /ih/ /k/ /ax/ /l/ /iy/ Sympathetically |
| 6     | /ow/ /l/ /z/ /iy/ /ax/ /m/ — — — Symposium |
| 7     | /t/ /ax/ /m/ — — — — — Symptom |
| 8     | B /ax/ /l/ /ih/ — — — — — Symbolic |
| 9     | — — — — — Symbolism |
| 10    | /iy/ — — — — — — — Symbolize |
| 11    | /el/ /l/ /m/ — — — — — Symbols |
| 12    | Ax — — — — — — — — Simmer |
| 13    | Eh /n/ /t/ — — — — — — Cement |

The oscillation of Lagrange polynomial is addressed as Runge (1901) phenomenon. This degrading effect can be well eliminated by repositioning sample’s node {xi} and interval modification.

To apply the proposed procedure, starting with the basic function modification in Equation 9 and rewriting it we have

$$l_i(x) = \frac{l(x) w_i}{(x-x_i)}, \quad (10)$$

where \(l(x)\) is

$$l(x) = \prod_{j=0}^{N-1} (x-x_j). \quad (11)$$

In parallel, the Barycentric weight function \(w_i(x)\) is

$$w_i(x) = \frac{1}{\prod_{k \neq i}^{N-1} (x-x_k)}, i = \{0,1,2,\ldots,(N-1)\}. \quad (12)$$

Substituting Equation 10 in Equation 8 gives

$$L(x) = l(x) \sum_{i=0}^{N-1} f_i \frac{w_i}{(x-x_i)}. \quad (13)$$

If Equation 13 is used for interpolating constant amplitudes equal to 1, then the resulted expression is

$$\sum_{i=0}^{N-1} \frac{w_i}{(x-x_i)} = 1. \quad (14)$$

Finally, the Barycentric Lagrange interpolation (BLI) can be formulated by substituting Equation 14 into Equation 13 which leads to

$$L(x) = l(x) \sum_{i=0}^{N-1} f_i \frac{w_i}{(x-x_i)}. \quad (13)$$
This modification tackles the problem of destructive oscillations on the interval boundaries by a transformation to another domain. The transformation is possible by Chebyshev points of the second kind as

\[ x_{i, ch} = \cos \left( \frac{i \pi}{(N-1)} \right), \quad i = \{0, 1, 2, \ldots, (N-1)\}. \]  

Equation 16 can be thought as a mapping scheme that transfers a uniformly spanned interval \( x_i \) to another interval called Chebyshev interval \( x_{i, ch} \).

The overall representations of the visual words, which are separated according to the three visual features, are shown in Figure 6: the upper (a), lower (b), and corner (c) visual speech signals.

Regarding to the speech processing field, it is possible to select the appropriate viseme table that has been selected as the visual domain representation reference. Selecting appropriate phoneme–viseme mapping table for analysis of phonemic structure would be beneficial in deriving the practical visual speech signal expressions that are adaptable to audio-visual speech recognition systems. In other words, the method of driving visual speech signal signature could be applied to the speech recognizer system as its collaborating visual domain. The mapping scheme, which is represented in Table 3, is a sample presented that can be applied based on the manner of articulation.

The consonants consist of eight subgroups of phonemes that have visual similarities during articulation, whereas vowels categorized into four subgroups. Although the English alphabet consists of 26 letters, there are 43 phonemes. This fact shows the difference between possible sound in words and the

Figure 6. The (a) upper, (b) lower, and (c) corner visual speech sample sets after adjusting the end points.
letters. Now it was time to find the transcribed words text that was marked according to this phonemic table (Table 3.).

**Table 3.** The Phoneme-Viseme Table (Potamianos, Neti, Luettin, & Matthews, 2004), Phonemes are Grouped Into Viseme Classes.

| Consonants | Vowels     |
|------------|------------|
| /p, b, m/  | /ih, iy, ix, ax/ |
| /f, v/     | /ae, eh, ey, ay/ |
| /l, el, r, y/ | /uw, uh, ow/ |
| /s, z/     | /ao, ah, aa, er, oy, aw, hh/ |
| /sh, zh, ch, jh/ |             |
| /t, d, n, en/ |             |
| /th, dh/   |             |
| /ng, k, g, w/ |             |

Results and Discussion

In Figure 6a to 6c, the visual speech signal $V_{BLI}^S w_n(f_r)$, $V_{BLI}^L w_n(f_r)$, and $V_{BLI}^C w_n(f_r)$, $m = 1, 2, 3, \ldots, 13$, are represented. The visual speech signals are processed for allocation of mathematical expressions to words. The words are categorized based on the hierarchical relations in phoneme sequences. A word is represented in visual domain with specific mathematical expressions. The mathematical expression of a visual word consists of three visual speech signals. These visual speech signals will be interpolated from three visual features located on the upper, lower, and corner of outer lip contour. The mathematical expressions were derived by the BLI method. The number of visual speech signals is set to three as the three visual features include the necessary information for lip dimension during articulation. The reason for using BLI is the ability of formulating the visual speech signals which involve the sample sets extracted from the visual feature points without Rounge effect.

The mathematical expressions of the visual speech data obtained from the upper, lower, and corner visual speech sample sets were called the visual speech signals. As it was mentioned earlier, the extracted samples from a visual feature were called visual speech sample set. Collection of three visual speech samples sets corresponding to the upper, lower, and corner visual feature points are representing a word. The first observation from Figure 6 suggests a coherence incremental pattern. It is shared with upper and lower visual speech sample sets for $f_r \in [0; 10]$ or from 0 to 379 milliseconds. However, the corner visual speech sample sets show more steady incremental changes. The second segment starts in 12th frame until the 25th frame where almost all visual sample
sets are obtained by upper, lower, and corner visual features reach to their minimum. From the 25th to the 34th frame, the trends of visual speech sample sets in all three cases are again increasing. Afterward, from 35th frame until the ending frame in each sample set, the trends are decreasing. The visual speech samples are sharing similarities in a portion of beginning frames. This is because the words’ phonemic structure has the same sequence of three phonemes (/s/+/ ih/+/m/+ . . . ).

Conclusion
In this article, a new method for mathematical formulation of visual speech information has been suggested. Since the concept of viseme does not provide a unique representation of visual phoneme, in addition to the unavailability of an explicit representation of the dynamics of lips during articulation of the words. The lips’ movements during articulating a specifically designed set of words were mathematically formulated in this work. The visual information is corresponding to a set of words which have the same sequence of triphones in their initial parts. The words are extracted from TIMIT database with hierarchical analysis. The mathematical expressions of the visual speech obtained from the visual features were called the visual speech signals. Normalized rational relations of visual speech sample sets were chosen for providing more compact versions of visual speech signals as well as preserving their scaling ability. Assigning a mathematical unique expression to visual speech information is a novel method to create a more precise database for words in comparison with the existing methods.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research and/or authorship of this article.

References
Adjoudani, A., & Benoit, C. (1996). On the integration of auditory and visual parameters in an HMM-based ASR. Berlin, Germany: Springer.
Alothmany, N., Boston, R., Li, C., Shaiman, S., & Durrant, J. (2010, September 15-17). Classification of visemes using visual cues. Proceedings of the International Symposium ELMAR, Zadar, Croatia.
Ananthakrishnan, G., & Engwall, O. (2008, December 8-12). Important regions in the articulator trajectory. Proceedings of the 8th International Seminar on Speech Production, Strasbourg, France.
Birkholz, P., Kröger, B. J., & Neuschaefer-Rube, C. (2011). Model-based reproduction of articulatory trajectories for consonant-vowel sequences. IEEE Transactions on Audio, Speech & Language Processing, 19, 1422-1433.

Bregler, C., Covell, M., & Slaney, M. (1997). Video rewrite: Driving visual speech with audio. In Proceedings of the 24th annual conference on SIGGRAPH 97 (pp. 353-360). New York, NY: Association for Computing Machinery.
Bregler, C., & Konig, Y. (1994, April 19-22). Eigenlips for robust speech recognition. IEEE International Conference on Acoustic Speech and Signal Processing, Adelaide, Australia.
Chin, S. W., Seng, K. P., & Ang, L.-M. (2012). Audio-visual speech processing for human computer interaction. In T. Gulrez & A. E. Hassanien (Eds.), Advances in robotics and virtual reality (pp. 135-165). Berlin, Germany: Springer.
Chiu, G., & Hwang, J. N. (1997). Lipreading from color video. IEEE Transactions on Image Processing, 6, 1192-1195.
Cohen, M., & Massaro, D. (1993). Modelling coarticulation in synthetic visual speech. In N. M. Thalman & D. Thalman (Eds.), Models and techniques in computer animation (pp. 139-156). Tokyo, Japan: Springer-Verlag.
Conrey, B. L., & Pisoni, D. B. (2003, September 4-7). Audiostream visual assynchrony detection for speech and nonspeech signals. Proceedings of the ISCA Tutorial and Research Workshop, St. Jorioz, France.
Cootes, T. F., Edwards, G. J., & Taylor, C. J. (1998, June 2-6). Active appearance models. European Conference on Computer, Freiburg, Germany.
Cox, S., Harvey, R., & Newman, J. (2008). The challenge of multispeaker lip-reading. Queensland, Australia: International Conference on Auditory-Visual Speech Processing.
Czup, L. (2000, October 16-20). Lip representation by image ellipse. International Conference on Spoken Language Processing, Beijing, China.
De Martino, J. M., Magalhaes, L. P., & Violato, F. (2006). Facial animation based on context-dependent visemes. Journal of Computers and Graphics, 30, 971-980.
de Vega, M., Alvarez, C., & Carreiras, M. (1992). Estudio estadístico de la ortografía castellana: La frecuencia silábica. Cognitiva, 4(1), 75-114.
Erber, N. P., Sachs, R. M., & DeFilippo, C. L. (1979). Optical synthesis of articulatory images for lipreading evaluation and instruction. Advances in Prosthetic Devices for the Deaf: A Technical Workshop, Rochester, NY.
Ezzat, T., & Poggio, T. (1998, June 8-10). MikeTalk: A talking facial display based on morphing visemes. Proceeding of Computer Animation Conference, Philadelphia, PA.
Fisher, C. G. (1968). Confusions among visually perceived consonants. Journal of Speech and Hearing Research, 11, 796-804.
Goldschchen, A. J. (1993). Continuous automatic speech recognition by lipreading (Doctoral dissertation). George Washington University, Washington, DC.
Goldschchen, A. J., Garcia, O. N., & Petajan, E. (1994). Continuous optical automatic speech recognition by lipreading. Pacific Grove, CA: IEEE Computer Society Press.
Hazen, T. J., Saenko, K., & Glass, J. R. (2004). A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments. New York, NY: Association for Computing Machinery.
Heracleous, P., Tran, V. A., Nagai, T., & Shikano, K. (2010). Analysis and recognition of NAM speech using HMM distances and visual information. IEEE Transactions on Audio, Speech, and Language Processing, 18, 1528-1538.
Huang, J., & Povey, D. (2005). Discriminatively trained features using fMPE for multi-stream audio-visual speech recognition.
Sadaghiani et al.

In INTERSPEECH-2005 (pp. 777-780). Retrieved from http://www.isca-speech.org/archive/interspeech_2005/i05_0777.html

Jiang, J., Alwan, A., Auer, E. T., & Bernstein, L. E. (2001). Predicting visual consonant perception from physical measures. Retrieved from http://www.seas.ucla.edu/spapl/paper/jiang_europospech01.pdf

Kim, J., & Davis, C. (2003, September 4-7). Testing the cuing hypothesis for the AV speech detection. International Conference on Audio-Visual Speech Processing, St Jorioz, France.

Kosmopoulos, D., & Chatzis, S. P. (2010). Robust visual behavior recognition. In IEEE Signal Processing Magazine (pp. 34-45). New York, NY: Institute of Electrical and Electronics Engineers.

Lee, J. S., & Park, C. H. (2006). Training hidden Markov models by hybrid simulated annealing for visual speech recognition. Taipei, Taiwan: IEEE.

Lucero, J. C. (2002). Identifying a differential equation for lip motion. Medical Engineering & Physics, 24, 521-528.

Massaro, D. W., Beskow, J., Cohen, M. M., Fry, C. L., & Rodriguez, T. (1999, August 7-9). Picture my voice: Audio to visual speech synthesis using artificial neural networks. International Conference on Auditory-Visual Speech Processing, Santa Cruz, CA.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264, 746-748.

Melenchón, J., Martínez, E., De La Torre, F., & Montero, J. A. (2009). Emphatic visual speech synthesis. IEEE Transactions on Audio, Speech, and Language Processing, 17, 459-468.

Montgomery, A. A. (1980). Development of a model for generating synthetic animated lip shapes. Journal of the Acoustical Society of America, 68(S1), S58.

Mortenson, M. E. (1997). B-spline curves. In M. Spencer (Ed.), Geometric modeling (pp. 113-142). New York, NY: Wiley.

Petajan, E. D. (1984). Automatic lipreading to enhance speech recognition. Atlanta, GA: IEEE Global Telecommunications.

Potamianos, G., & Graf, H. P. (1998, May 12-15). Discriminative training of HMM stream exponents for audio-visual speech recognition. International Conference on Acoustics, Speech and Signal Processing, Seattle, WA.

Potamianos, G., Neti, C., Luettin, J., & Matthews, I. (2004). Audio-visual automatic speech recognition. An overview. Cambridge, MA: MIT Press.

Revéret, L., & Benoît, C. (1998). A new 3D lip model for analysis and synthesis of lip motion in speech production. Terrigal, Australia: Auditory-Visual Speech Processing.

Rogozan, A., & Deléglise, P. P. (1998). Adaptive fusion of acoustic and visual sources for automatic speech recognition. Speech Communication, 26, 149-161.

Runge, C. (1901). Über empirische Funktionen und die Interpolation zwischen aquidistanten Ordinaten. Zeitschrift für Mathematik und Physik, 224-243.

Saenko, K., Livescu, K., Glass, J., & Darrell, T. (2009). Multistream articulatory feature-based models for visual speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31, 1700-1707.

Su, Q., & Silsbee, P. L. (1996). Robust audiovisual integration using semicontinuous hidden Markov models. Philadelphia, PA: International Science Congress Association.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America, 26, 212-215.

Teissier, P., Robert-Ribes, J., & Schwartz, J. L. (1999). Comparing models for audiovisual fusion in a noisy-vowel recognition task. IEEE Transactions on Speech and Audio Processing, 7, 629-642.

Williams, J., Rutledge, J., Garstecki, D., & Katsaggelos, A. (1997, June 23-25). Frame rate and viseme analysis for multimedia applications. Proceedings of IEEE Multimedia and Signal Processing Conference, Princeton, NJ.

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

Author Biographies

Mohammad Hossein Sadaghiani has received his PhD in the University of Nottingham, Malaysia Campus, and still continuing his research on voice recognition.

Niusha Shafiabady is currently a faculty member at the University of Nottingham, Malaysia Campus. Her research interests are applications of artificial intelligence in different areas of science.

Dino Isa is the professor of intelligent systems at the University of Nottingham, Malaysia Campus. His research interest is application of artificial intelligence in different areas of science and technology.