Visual Speech Recognition using Neural Networks

1

Dewey L. Sia, Ateneo de Davao University   
Zuleika a. yee, Ateneo de Davao University

Lip reading is a notoriously difficult task for humans, even for the hearing-impaired, who achieved roughly 17±12% lip reading performance for monosyllabic words. Driven by the recent rise of deep learning techniques, machine lip reading or visual speech recognition research has started to surface and such systems have started to outperform humans. Lip reading systems can be used for improving hearing aids, silent dictation in public places, biometric identification, audio-visual speech recognition, as well as graphics and animation, among others. There is already a number of researches done on lip reading English words, with the state of the art system, LipNet, achieving a 95.2% sentence-level lip reading accuracy on seen speakers. This research will primarily focus on applying deep learning techniques to build a neural network capable of lip reading Filipino words. The Filipino vocabulary will consist of 24 words with different syllabications based on the six syllabication rules provided by the KWF. The data used for training and evaluating the model is collected locally by the researchers in a semi-controlled environment. The raw data will be in the form of video, which is then preprocessed into images or video frames. The preprocessed data set will consist cropped images of lip regions of various speakers. Computer vision techniques such as face and lip localization are used to process the dataset. The neural network model will be built on top of preexisting models that competed in the ImageNet Large Scale Visual Recognition Competition (ILSVRC), data augmentation and model fine-tuning are used in an effort to improve classification accuracy. The implementation of the neural network model will be done primarily in Keras, using the TensorFlow backend.

General Terms: Long-Short Term Memory, Convolutional Neural Networks, Transfer Learning, Early Stopping

Additional Key Words and Phrases: Phonemes, Visemes, Data Augmentation

# INTRODUCTION

## Background of the Study

People who are deaf, mute, or both, have a difficult time interacting with others daily. Deaf people cannot hold conversations well because they are unable to decipher what the other party is saying. Mute or speech impaired people have a hard time communicating because they cannot speak or do not speak well. Aids of different kinds are employed to help make communication for these people easier. Hearing aids, pen and paper, and sign language are examples of these. In these situations, some people may attempt lipreading the other party in order to aid comprehension. However, even for the mute, their lip reading accuracy is considered very poor (17±12%) [2].

With the recent rise of deep learning, lipreading tasks have become easier for machines, which have already trumped human lipreading accuracy. Machine lipreading is an emerging field of study, which have many applications aside from assisting the hearing-impaired such as speech recognition in noisy environment, which are still challenging for audio speech recognition approaches. Many studies have tackled the lipreading problem using different machine learning and deep learning approaches to different effectiveness, but none of them were done on Filipino lipreading [2, 17, 27, 31, 48, 57, 58]. It may be appropriate to apply machine lipreading within the Philippines if the model were trained on Filipino speakers. The input data will be a sequence of video frames extracted from video clips of speakers, which is an approach only few papers applied [31, 51]. The proponents explore the challenges of visual speech recognition with the Filipino languages.

## Problem Statement

## Lipreading is a notoriously difficult task for humans, even for the experienced. This study aims to provide a way for computers to lipread people by building a deep learning model. Machine lipreading can contribute to the speech and hearing-impaired community. In particular, the researchers build a neural network model which is fed video frames and corpus data. The model will then be used to predict tokens within unfamiliar videos. Research will be conducted on the following:

1. What methods and techniques can be used to improve lipreading accuracy given the data?
2. How does our approach perform compared to other baseline approaches?
3. How does predicting Filipino words/phrases affect the performance of the neural network model?
4. Can our model perform fast enough to make real time predictions? How much delay due to processing is acceptable?
5. How do we effectively extract and segment lips and areas of interest from the video frame data?

## Objectives

This study aims to improve lip reading accuracy across different familiar and unfamiliar people. Lipreading can be applied to aid in many tasks. This study will have a focus on gauging performance between machine and human lipreaders, especially the hearing and speech impaired. The model used in predicting speech should preferably also be quick to produce output in order to lipread in real-time.

The study has the following specific objectives:

1. To determine methods and techniques that improve lipreading accuracy.
2. To be able to compare our approach from other baseline approaches.
3. To know how can the lipreading Filipino words/phrases affect the performance of the neural network model.
4. To determine if our model can perform fast enough to make real time predictions and how much processing delay is acceptable.
5. To know how extracting and segmenting lips and areas of interest from videos can be done effectively.

## Significance of the Study

Deaf people use sign language or bring paper and pen to communicate with other people. Partly deaf people sometimes lipread for them to acquire a more accurate interpretation of what other people say. In both these areas, machine lipreading will be able to aid its user. Machine lipreading has various other applications aside from assisting the mute and deaf. Some other applications of machine lipreading include:

* Improved hearing aids
* Silent dictation in public spaces
* Biometric identification
* Silent-movie processing
* Security
* Audio-visual speech recognition
* Graphics and animation
* Aid for persons with speech impairments (e.g. laryngectomees, whose voice box (larynx) has been removed)

## Scope and Limitations

# The neural network model will handle only with lipreading the frontal view of the subjects’ faces. All data used for training the model are gathered locally by the researchers. Research will focus on doing visual-only speech recognition and will not include a feature fusion of this and audio speech recognition. The language the model is trained to predict is Filipino, primarily because there is a distinct lack of research regarding visual speech recognition in Filipino. The richness of the model’s vocabulary will depend heavily on the data used to train it, thus it should not be expected to predict tokens it was not trained on. Similarly, the researchers do not promise the model to be able to do real time predictions but will look into it.

Variants of lip reading research deal with a combination of lip reading levels, controlled and uncontrolled environments, and vocabulary size. Chung et al. (2016), along with Google DeepMind, trained a convolutional neural network that predicts speech at the word level using a dataset of videos from the wild. They further expanded on their research by incorporating audio-visual speech recognition. Another research by Assael et al. (2017) introduced a model trained on the GRID corpus, the largest publicly available lipreading dataset. They used convolutional neural networks and long-short term memory to produce a model that predicts speech at the sentence-level in a controlled environment and achieved a state of the art accuracy of 95.2% on recognized speakers. However, the nature of their dataset is such that it has very limited vocabulary. This research will similarly use neural networks as the deep learning model. The researchers aim to provide model that can predict speech on Filipino speakers.

# REVIEW OR RELATED LITERATURE

## Lipreading

Lipreading is extremely difficult for humans, even more so when lacking context. Traditionally, lipreading approaches separate the problem into two stages: one being designing or learning visual features, and the other being prediction [17]. Human lipreading performance is poor. Hearing-impaired people achieve an accuracy of only 17±12% (even for a limited set of 30 monosyllabic words) and 21±11% (30 compound words). The current challenges of machine lipreading are the speed at which a person speaks and the video resolution.

### Phoneme

Phoneme are the smallest unit of speech that distinguishes one word from another. For example, the word element p in “tap” distinguishes it from the words “tab”, “tag”, and “tan.” [2].

#### Allophone

An allophone is a variant of phoneme. Speakers of a language usually have difficulties differentiating one allophone from another because differences of allophones of the same phoneme do not bear significance for these are not used to distinguish one word from another. For example, the t sounds in the words “hit,” “tip,” and “little” are allophones; phonetically considered to be the same sound, although they are phonetically different in terms of aspiration, voicing, and point of articulation [1].

#### Phoneme Classification

There has been a number of attempts to automate lipreading in the recent years. Approaches that do not use deep learning require heavy preprocessing of frames to extract image features, temporal preprocessing of frames to extract video features or other types of handcrafted vision pipelines (Matthews et al., 2002; Zhao et al., 2009; Gurban & Thiran, 2009; Papandreou et al., 2007; 2009; Pitsikalis et al., 2006; Lucey & Sridharan, 2006; Papandreou et al., 2009)[2]. For approaches that apply deep learning, most perform only word or phoneme level classification. LipNet (Yannis et al., 2016) produced the first end-to-end sentence-level lipreading system.

### Viseme

Viseme is the visual equivalent of a phoneme. A working equivalent of visemes is that it is a set of phonemes that have identical appearance on the lips [8]. As of current time, there is no proven function between or relation (often presented as a map) between visemes and phonemes. Current knowledge on the topic is limited. Visemes are also called visual phonemes (Fisher, 1968). According to Fisher (1968), some visemes are more commonly confused by people when lipreading, specifically, there are 23 initial consonant phonemes which are mostly metrically confused, and observations were similar for final consonant phonemes [24].

### Phoneme-to-viseme maps

A critical assumption of many lipreading machine systems is that visemes can be mapped to the units of speech, the phonemes. There are more phonemes than visemes, this results in a many-to-one relationship or phoneme-to-viseme map [4]. Although quite a number of viseme-phoneme maps have been published, their effectiveness have not been tested, particularly on visual-only lipreading (many works use them for audio-visual speech). A phoneme may fall into one viseme class but a viseme may map to many phonemes (Many-to-one). Most phoneme-to-viseme (P2V) mappings are consonant-only mappings, which are derived from single speaker data. , Bear, et al. (2017) consider a combination of various known mappings (15 consonant maps and 8 vowel maps) which are paired with each other to produce 120 P2V maps [8]. Division of phonemes across viseme classes will vary with each map. Silent phonemes are usually omitted.

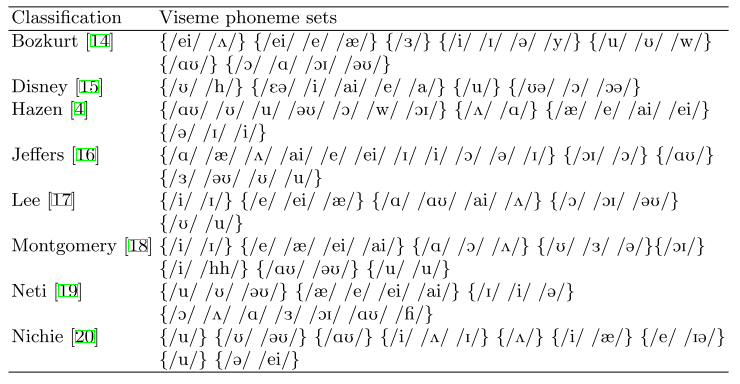


Fig. 1. Bear, et al. (2017)’s Vowel Viseme-to-Phoneme (British phonetics) maps from Which phoneme-to-viseme maps best improve visual-only computer lip-reading?

P2V mappings are contractive. Each mapping varies in the number of viseme classes. Thus, in Fig 2. below, the Woodward map covers 24 consonant phonemes to 4 visemes with a confusion factor of 4/24 (0.167), whereas the Jeffers map covers 23 vowel phonemes which are mapped to 8 visemes.

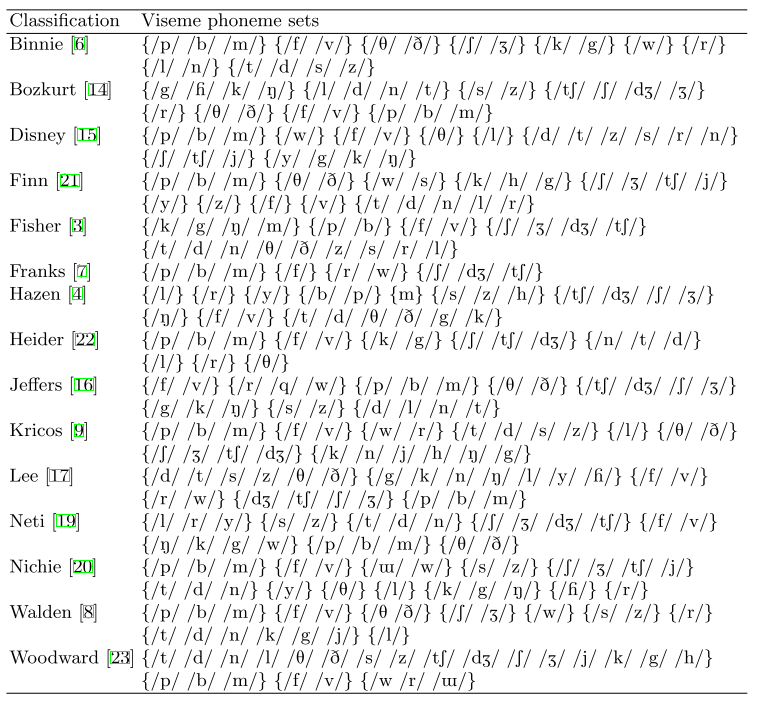


Fig. 2. Bear, et al. (2017)’s Consonant Viseme-to-Phoneme (British phonetics) maps from Which phoneme-to-viseme maps best improve visual-only computer lip-reading?

In the case of Bear, et al., they deliberately omitted several phonemes from the mappings due to the their phonetic data set being British, while some mappings are American diacritics. The Filipino phonetic set also has a different P2V mapping. However, no literature were found discussing Filipino English mappings.

It is possible that the Filipino phonetic set also face the diacritic differences when using non-local English phonetic sets. No literature could be found addressing this issue.

#### Filipino Phonology

Distinctive features of the Tagalog language are lexical stress and glottalization. Lexical stress being the stress or emphasis given to a certain syllable in a word. Glottalization is defined as either the complete or partial closure of the glottis during articulation. Tagalog words are often distinguished from one another by the position of the stress and the presence of the glottal stop [56].

Filipino English speakers have a certain accent which is often referred to as “Filipino Accent”, which can be difficult for native English speakers to understand. Filipinos often mix up several consonants and vowels in English words they speak.

Philippine English is a rhotic accent. Therefore, /r/ phonemes are pronounced in all positions. For non-native Filipino English speakers, Philippine English phonological features are heavily dependent on their mother tongue, although other factors such as the colonization of the Spanish also influenced how many Filipino pronounce English words. The most distinguishable feature of Philippine English is its lack of fricative consonants, particularly /f/, /v/ and /z/. From this, it can be seen that Filipino English has different phoneme-viseme mappings from the ones presented thus far [3].

##### Filipino Language

The Filipino language is considered to be syllable-timed, which means that Filipino words are observed to have almost the same syllable lengths regardless of the number of stresses in its construction in statements [30]. The Komisyon ng Wikang Filipino (KWF), in its document entitled Ortograpiyang Pambansa (2013), recommended a set of rules in syllabicating Filipino words (See Fig. 3.) [47].

|  |  |  |
| --- | --- | --- |
|  | **RULE** | **EXAMPLES** |
| 1 | A syllable may be comprised of one vowel alone but a consonant needs a vowel to make a syllable. Syllables with two or more letters can have only one vowel. However, a syllable can have one or more consonants. | ako (a-ko)  sila (si-la)  ulam (u-lam)  baul (ba-ul)  almusal (al-mu-sal) |
| 2 | If there are two or more consecutive vowels in a word, the vowels are separated into different syllables. These consecutive vowels may appear at the beginning, in the middle, or at the end of a word. | aagaw (a·a·gaw)  uulan (u·u·lan)  alaala (a·la·a·la)  asotea (a·so·te·a)  baunan (ba·u·nan) |
| 3 | If there are two consecutive consonants in a word, the first consonant joins the vowel before it in one syllable. The second consonant becomes part of the next syllable. | ambon (am·bon)  pinggan (ping·gan)  madre (mad·re)  tigre (tig·re)  serbisyo (ser·bis·yo) |
| 4 | If there are three different consecutive consonants somewhere in the middle of a word, the first two consonants join the vowel before them in one syllable. The third consonant becomes part of the next syllable. | breyslet (breys·let)  transportasyon (trans·por·tas·yon)  inspirasyon (ins·pi·ras·yon)  ekskursiyon (eks·kur·si·yon)  eksperimento (eks·pe·ri·men·to) |
| 5 | If there are three consecutive consonants somewhere in the middle of a word, and the first consonant is m or n and the next two consonants are any of the consonant clusters bl, br, dr, pl, or tr, then the m or n joins the vowel before it in one syllable and the consonant cluster becomes part of the next syllable. | sum·bre·ro (sum·bre·ro)  miyembro (mi·yem·bro)  balandra (ba·lan·dra)  timpla (tim·pla)  simple (sim·ple) |
| 6 | If there are four consecutive consonants somewhere in the middle of a word, the first two consonants join the vowel before them in a syllable. The last two consonants become part of the next syllable. | abstrak (abs·trak)  ekstra (eks·tra)  ekspres (eks·pres)  eksklusibo (eks·klu·si·bo)  eksplorasyon (eks·plo·ras·yon) |

Fig. 3. Rules of syllabicating Filipino Rules recommended by the Komisyon ng Wikang Filipino

##### Filipino and English Comparison

On the other hand, the English language is known to be a stress-timed language [54]. This means that the time length of pronouncing syllables from a word may vary. Stressed syllables have regular intervals, while unstressed syllables shorten to fit in this rhythm [46].

##### Philippine Consonants

According to Bautisa et al. (2008), some consonant changes that apply to most non-native speakers of English are:

* The rhotic consonant /r/ varies between a trill [r], a flap [r], and an approximant [ɹ].
* The fricative /f/ is replaced by stop consonants [p] while /v/ is replaced by [b].

##### Philippine Vowels

Bautista et al. (2008) states that Philippine English vowels are pronounced according to the letter they symbolize. The vowels ⟨a, e, i, o, u⟩ are generally pronounced as [a, ɛ, i, o, u], respectively. Words ending in -ple, -fle, or -ble (apple, waffle, humble) are pronounced with an [ol].

#### New Phoneme-to-Viseme Mappings

For the existing P2V mappings, Lee provides the best consonant and vowel maps. Bear et al. (2017) explores the possibility of alternatives that perform better. They propose two alternatives:

* Tightly confused maps, where all phonemes within each viseme can be confused with each other in the phoneme recognition. Some viseme sets will contain single-phoneme visemes
* Loosely confused maps, where, from the tightly confused mapping, phonemes which are not confused with all other phonemes in a viseme set are merged with the viseme class with the largest confusion.

For each type of mapping, there are 2 sets created with vowel and consonant phonemes mixed and separate.

Given their limited data set, Bear et al. concluded that:

*Mixed consonant and vowel maps perform better than split consonant and vowel maps.*

Comparing loosely confused maps and tightly confused maps, tight confusions perform better for 2 out of 4 speakers and equal for a third. As shown in Fig. 5, speaker-dependent, loosely coupled, split viseme sets perform the best for word classification.

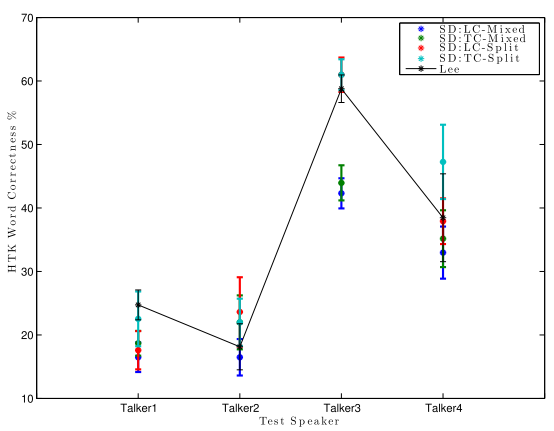


Fig. 4. HTK word correctness based on phoneme recognition confusions. SD = Speaker Dependent, LC = Loosely coupled, TC = tightly coupled, Mixed = mixed vowels and consonants within viseme classes, Split = separated vowel and consonant visemes. (Bear et al, 2017)

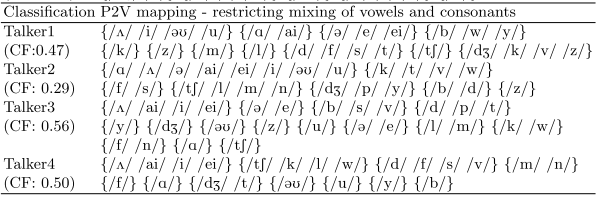


Fig. 5. Speaker-dependent, loosely coupled, split viseme mapping

### Recognition Method

Measuring correct viseme-to-phoneme translation involves a viseme transcription produced by converting phonetic transcript of the training data to viseme labels, then testing the transcription through a mapping.

### Lipreading Levels

#### Digits

Fu et al. (2008) used the AVICAR dataset consisting 851 utterances of digits to build a lipreading system. They presented a new pattern classification algorithm called Locality Discriminant Graph and developed a novel framework to successfully apply it to. Their system was able to obtain an accuracy of 37.9% [26]. Papandreou et al. (2009) pursued lipreading digits by using the CUAVE dataset which consists of 1800 digit utterances, and achieved an accuracy of 83.0% [2].

#### Words and Phrases

Chung and Zisserman (2016) introduced the ‘Watch, Listen, Attend and Spell’ network model that can transcribe speech into characters and used the word-level BBC TV dataset to predict phrases and sentences. Although their dataset includes sentences, they are only limited to predicting sentences character-by-character and word-by-word. Hence, they do not cover sentence-level sequence prediction. They claim an accuracy of up to 65.4% [17]. According to Hassanat et al. (2009), the representations of the words and phrases should capture specific visual information closely associated with the spoken word to allow for the recognition for the word/phrase and to distinguish it from others [33].

#### Sentences

LipNet, the state of the art end-to-end sentence-level lip reading software, integrated spatiotemporal convolutions, Long Short-Term Memory (LSTM) recurrent network, and Connectionist Temporal Classification (CTC) Loss into their implementation. The LSTM recurrent network is responsible for sentence-level sequence prediction and speaker independence. Hence, the need to segment videos into words before predicting a sentence was eliminated. LipNet claims a lipreading accuracy of up to 95.2%, better than professional lipreaders. However, the dataset they used is only limited to valid sentence patterns generated by four commands, four colors, four prepositions, twenty five letters, ten digits and four adverbs [2].

### Speaker Dependence

Speaker dependence defines how a machine learning algorithm learns visemes from different speakers or people with different viseme confusions, and determines the contents of a phoneme-to-viseme map [6]. Werda et al. (2007) provided an approach in recognizing French visemes by grouping them into three differentiable groups [62].

#### Speaker-Dependent Visemes

Speaker-dependent visemes are operated by learning the unique qualities of a single person’s voice, which is moderately similar to voice recognition. Hence, new users shall train a software themselves by reading a few pages before it becomes usable, in order for the software to analyse and learn the way they talk, and to achieve higher accuracy rates [55].

The mapping of speaker-dependent visemes is derived from the phoneme recognition confusions of an individual speaker. It enables the creation of a set of phoneme-to-viseme maps where each speaker has a unique set of visemes containing different number of phonemes, assuming that there are more than one speaker whose visemes are studied [6].

#### Speaker-Independent Visemes

Speaker-independent visemes can be used to recognise unfamiliar speakers whose footage was not used in model training. This type of viseme is essential for applications like Interactive Voice Response systems or real-time applications that require learning visemes. However, its generalization reflects lack in accuracy [55].

The phoneme-to-viseme map of speaker-independent visemes contain visemes that are based on the phoneme confusions of other speakers for each speaker. Unlike the mapping of speaker-dependent visemes, the number of visemes are similar in quantity [6].

## Approaches/Architectures

### Deep Learning

Deep learning, also known as deep structured learning, is a subfield of machine learning methods based on learning data representation [15]. Deep learning architectures including deep neural networks, deep belief networks, and recurrent neural networks, are now capable of aiding in the fields of computer vision, speech recognition, natural language processing, audio recognition, and the such. Deep learning have often produced results superior to that of human experts [41]. Many recent approaches in visual speech recognition employ deep learning and neural networks to excellent results.

#### Convolutional Neural Networks

Assael et al. (2016) developed an end-to-end sentence-level sequence lipreading model called LipNet, which maps variable-length video sequence frames to text using spatiotemporal convolutions, a recurrent network, and the connectionist temporal classification loss. The model was trained on the GRID corpus (contains audio and video recordings of 34 speakers with each speaker producing 100 sentences each; a total of 28 hours of recording) [2].

#### Spatiotemporal Convolutions.

A convolutional neural network CNN that contains stacked convolutions operating spatially over an image. This advances performance in computer vision tasks such as object recognition using images [2].

#### Long Short-Term Memory

Wand et al. (2016) used an LSTM neural network model which yielded better accuracy than conventional approaches [57]. After being tested on the validation set speakers against a baseline support vector machine (SVM) classifier using conventional features, coupled with Eigenlips and Histograms of Oriented Gradients (HOG), the LSTM model achieved 79.6% accuracy on the GRID corpus compared to Eigenlips + SVM model accuracy of 70.6% and HOG + SVM model accuracy of 71.3% [57].

Gutierrez et al. (2016) used 4 model architectures for their machine lipreading research on the MIRACL-V1 dataset: (1) CNN + LSTM baseline, (2) Deep CNN + LSTM, (3) Frozen VGG + LSTM, (4) Fine-tuned VGG + LSTM. They achieved a relatively good accuracy on the dataset for seen speakers, with the highest test accuracy being 59% from the fine-turned VGG + LSTM model [31]. LSTM is preferred over Recurrent Neural Networks

Garg et al. (2016) approached the classification problem on MIRACL-V1 using two approaches:

1. Per data instance, each image in a sequence to form one larger image.
2. Use Long-Short Term Memory (LSTM) layers to handle sequence features and infer temporal information so the CNN layers do not need to.

Due to the relatively small size of the dataset, a pre-trained model (VGG) was used for convergence without overfitting. Long-Short Term Memory (LSTM) layers were chosen over regular Recurrent Neural Networks (RNN) because LSTMs do not face the vanishing gradient problem [27].

#### Connectionist Temporal Classification Loss

Widely used in modern speech recognition. Eliminates the need for training data that has inputs that align with target outputs a neural network output and scoring function. This is used for training recurrent neural networks provided variable timing length inputs. We refer to the task of labelling of sequences of unsegmented data as temporal classification [29]. LipNet is trained end-to-end using CTC and thus does not require alignment [2].

### Non-deep learning approaches

#### Hidden Markov Model

Using the Hidden Markov Toolkit (HTK), Bear et al. (2017) built viseme-level hidden Markov model (HMM) recognizers with 5 states and 5 mixture components per state. Leave-one-out seven fold cross validation was used (seven folds since their data set, the AVL2, has 7 utterances of the alphabet per talker). Their HMMs are initialized using “flat start” training, and then re-estimated 8 times and afterwards force-aligned using HTK’s HVite. After re-estimating the HMMs 3 more times, the training is complete [7].

#### Support Vector Machines

In the study “A New Visual Speech Recognition Approach for RGB-D Cameras”, Rekik et al. (2014) proposes a relatively new approach for lipreading using 2D images and their corresponding depth data. The study used support vector machines (SVM) trained on the MIRACL-VC dataset which the researchers themselves collected. SVMs were used due to their ability to find a globally optimum decision function to separate different classes of data. They found out that the linear kernel for SVM performed the best with an accuracy of 79.2% on classifying speaker-dependent phrases and 63.1% on classifying speaker-independent words [51].

#### K-Nearest Neighbor

In the study “A Lip Reading Application on MS Kinect Camera”, Yargic et al. (2013) proposed a lipreading application for classifying Turkish color names using the K-Nearest Neighbor (KNN) classifier, where the input data is classified by using the distance to the nearest neighbor. During classification, each color class is assigned to the nearest neighbor based on the Euclidean and Manhattan distances. The data was gathered using an MS Kinect camera in a normally lit environment with 12 frames per second and the resolution of 1280x960. The Face Tracking SDK was used to detect face and extract lip points. The data is then preprocessed offline to segment the words to be classified using KNN [66].

### Data Preprocessing

#### Lip Localization

Lip localization is also known as lip segmentation. This is also the first step in the lipreading process. For accurate visual information extraction, a requirement is reliable mouth movements. Methods used for extracting lips may be semi-automatic or automatic. There is often a trade-off between accuracy and speed in lip detection [32].

##### Semi-automatic Lip Localization

The semi-automatic method of lip extracting requires the manual selection of a pixel point in the first frame to initiate the process. Semi-automatic lip tracking using geometric information. The drawback is the need for manual selection of initial seed for detection. The initialization stepped can be achieved by using the jumping snake algorithm proposed by Eveno et al. [20], which finds the upper lips boundary points. This is derived from the classical snake algorithm and improves upon its 2 deficiency on parameter selection and initial position selection [39]. Lalitha et al. (2016) traced the following video frames using the Kanade-Lucas optical flow algorithm [42].

##### Automatic Lip Localization

Werda et al. (2007) proposed an automatic lip feature extraction approach using geometric information which they called ALiFe. This approach uses the snake algorithm as the lip contour algorithm to segment the lips. After which, the points of interest (POI) are identified; these are the horizontal and vertical points of the lip contour. The detected POIs from the 1st frame is then traced on the following frames using the Template Matching method. The template matching method tracks a block of the following frame and determines candidate points. The candidate points undergo a voting method to find the nearest POI in the following point [62].

##### Hybrid model for Lip Localization

Mahdi et al. proposed an automatic hybrid model which uses both geometric and color information for tracking POIs on mouth [43]. Lip color is distinguished from skin color using HSV color model and rg chromacity. The saturation component is used to localize the mouth corners from the mouth region. The proposed hybrid solution can tolerate more noise and artefacts in the image compared to the 2 geometric-only localization methods above.

##### Model-based approaches

###### Snakes

Also called Active Contour Models, snakes minimize the energies of the spline iteratively to fit local minima [39]. A number of problems arise from using snakes. They can fit the wrong feature, such as the nose or the chin, especially if the initial position is far from the lips. Facial hair also affects its performance. Snakes may sometimes be difficult to tune parameters for and takes a long time to converge (several seconds) [32].

###### Active Shape Models

Active Shape models (ASMs) are statistical models of object shapes. A shape is represented as a set of labelled landmarks. ASMs are trained on a set of landmarked object in images and produces a global shape variation from the training set. PCA was used. According to Milborrow et al., ASMs perform better with more landmarks [44]. Kalman filters were used by Caplier to speed up convergence time [10].

###### Active Appearance Models

Active Appearance models (AAMs) are the same as ASM but extends upon using the edge profile along a landmark by including grey scale and shape information of the whole image. After 15 iterations, AAMs achieved good lip detections results. This is a generative model which during fitting of an object, aims to recover its parametric description through optimization. However, AMMs are not necessarily fast as there are more parameters to handle [19]. Bear et al. (2017) used an active appearance model to track talkers’ faces, and extracted common shapes and features for training. AAM features are known to outperform other feature methods in machine visual-only lipreading [5].

###### Deformable templates

Deformable templates are parameterized mathematical model used for object tracking. It can adapt itself to fit a given object. This is a useful shape model because of it is flexible and able to impose geometrical constraints on the shape and to integrate local image evidences [67]. The drawbacks of this method include high computational time complexity, unexpected template shrinkage, and template rotation [65].

##### Image-based approaches

###### RGB approach

In the RGB color space, the skin and lip pixels have different components. Red is dominant in both skin and lips, while there is more green than blue in skin color. Gomez et al. proposed a lip detection method where an image is transformed by a combination of red, green, and blue in the RGB color space. After which, a high pass filter is applied to highlight the details of the lips in the new image. Finally, both the high pass filtered image and transformed image is converted to a binary image. The largest area in the binary image is recognized as the lip area [28].

###### HSV (Hue Saturation Value) approach

HSV is closer to how humans describe and understand color. Hue represents the dominant color perceived by humans. Hue Is a key feature in lip detection because the hue value in lip pixels are smaller than that of the face pixels [42].

###### Hybrid Edge

A combination of color and luminance information. This uses hues to differentiate between the lips and skin color. The use of the pseudo hue with gradient vector of the intensity of the upper, middle, and lower lip sections allow a reliable estimation of several key point positions [21].

###### YCbCr approach

The YCbCr approach uses Particle Swarm Optimization to obtain a map based on the properties of the lip region which has high Cr and low Cb values. Jaroslav (2004) uses the fact that lips are seen as more red than other parts of the face to maximize the Cr and Cb components and the output image is thresholded to locate the mouth area [37].

###### Nearest Color approach

Hassanat et al (2010) remedied the above difficulties by incorporating color information into their scheme, despite making it vulnerable to light conditions. They proposed a “nearest color” method based on the YCbCr approach to find at least any part of the lip. The system is then trained on the pixel values of the part of the lip, clustering the pixels depending on the training data.

For their initialization step, they used the Lip-Map formula, which is designed to increase pixel values based on Cb and Cr. Because the Cr component is higher for the lips than for the face, the Lip-Map formula completes the idea that lip regions have high Cr and low Cb values. Although the Lip-Map formula cannot detect the whole lip region, the Hassanat et al. stipulates that it can, even at the worst case, a small part of the lip without being affected by other facial features such as facial hair. This results in it being more ideal than other approaches when doing the initialization step for lip detection.

Their “nearest color” method achieved an accuracy of 91%. This lip detection method is affected by variation in the skin and lip color due to ethnicity. In terms of ethnic groups, it achieved the lowest accuracy for Africans because the color of their skin and lips is very similar. This method is fast enough to be applied real time for online applications. However, the authors agreed that 91% lip localization accuracy is not enough for a robust lipreading system and requires further work [32].

### Facial Detection

Facial detection is the task of locating a face or faces in an image [23]. Unlike facial recognition, facial detection does not have to verify a person’s identity during processing [50]. However, the human face is even harder to detect than other objects. Analyzing images for facial detection is time consuming and difficult because of the different shapes and pigmentations of the human face [45][64].

Approaches in facial detection may depend on a scenario. It can be either in a controlled environment, colored images, or images in motion. A controlled environment is the simplest and most straightforward case for the detection of edges of a human face. Faces’ skin colors in colored images can be helpful in finding faces but can be weak, depending on light conditions. Images in motion or videos give the chance to use motion detection to localize faces [11].

Some approaches used in accurate facial detection are edge-orientation matching [25], Hausdorff Distance [36], Viola-Jones Algorithm [61], Histogram of Oriented Gradients (HOG) and deep learning.

#### Haar Cascade Classifiers

The Haar Classifier is an algorithm created by Paul Viola and Michael Jones, and improved by Rainer Leinhart. These classifiers are trained from multiple positive and negative images. Positive images are images that have faces on them, which is contrary to negative images [12].

It starts by extracting Haar-like features or windows from each image (See Fig. 6). Next, these windows are placed on the picture to come up with a single feature. This single feature is calculated by subtracting the sum of pixels under the white part of the window from the sum of the pixels under the black part of the window. Thus, all possible sizes and locations of each image are placed, in order to calculate plenty of features (See Fig. 7). However, because these features are mostly irrelevant, the irrelevant features are automatically disregarded using the Adaboost technique. Adaboost is a training process for face detection which is capable of disregarding features that are not useful in improving a classifier [22].

Using Haar cascade classifiers result to high detection accuracy and low false positive rate. However, this algorithm is computationally complex and slow, requires longer training time, is less accurate in black faces, has limitations in difficult lightening conditions, and is less robust to occlusion [22].

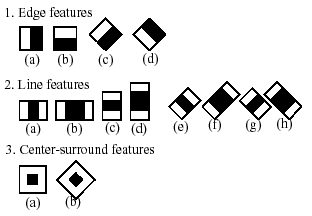


Fig. 6. Examples of Haar-like features or windows [12].

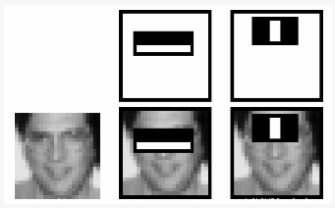


Fig. 7. Different stages in visualization [12].

#### Local Binary Patterns Cascade Classifiers

Local Binary Patterns, like any other classifier, also needs to be trained on hundreds of images. It is a visual or texture descriptor, which implies that it is efficient for face detection because our faces are also composed of micro visual patterns [22]. LBP features have been very useful in constructing applications such as texture classification, segmentation, image retrieval, and surface inspection [13].

LBP features are found by dividing each training image into blocks (See Fig. 8). For each block, the LBP looks at 9 pixels (3x3 window) at a time, with special focus on the center of the window. It compares the central pixel value to its neighboring pixel values. The value is set to 1 if the neighbor pixel is greater than or equal to the center pixel. Otherwise, its value is set to 0. Next, it forms a binary number in a clockwise order and converts the obtained binary number into a decimal number (See Fig. 9). The decimal number is now the new value of the center pixel. This is done for every pixel in the block. Finally, it converts each block values into a histogram (See Fig. 10) and concatenates these block histograms to form a feature vector for one image which contains all the needed features [22].

LBP cascade classifiers are computationally simple and fast, require shorter training time, and are robust to local illumination changes and occlusion. However, they are less accurate, having a high false positive rate [22].

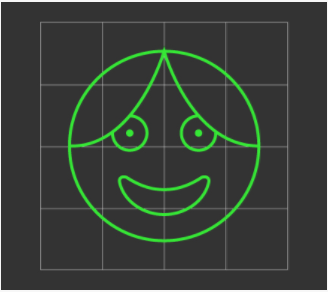


Fig. 8. An example of LBP Windows [22].

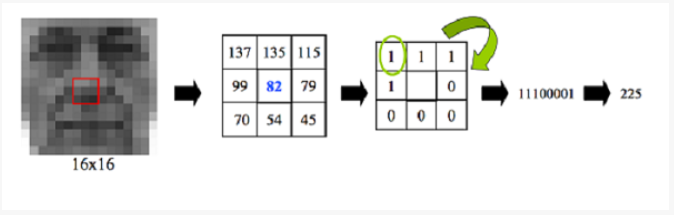


Fig. 9. An example of LBP conversion to binary [22].

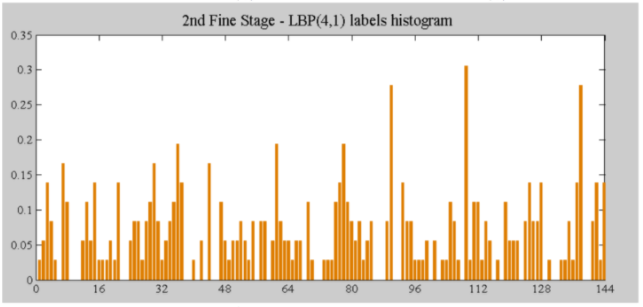


Fig. 10. An example of an LBP Histogram [22].

### Facial Landmarking

Facial landmarking, also known as facial feature detection, is the process of finding the location of different facial key points or the parts if the face. It has contributed to the development of face recognition, head pose estimation, face morphing, virtual makeover and face replacement. [23] Fig. 6 shows how researchers Chen et al. applied facial landmarking in an image. Specifically, the landmarks are divided into two categories namely key points and inserted points, which together allows the system to describe the face shape accurately [14].

Facial landmark tracking involves facial landmarking in videos or real-time. [9] Facial landmarking done in each frame is inefficient for dynamic footage, since the location information of facial features from different frames is not used [63].

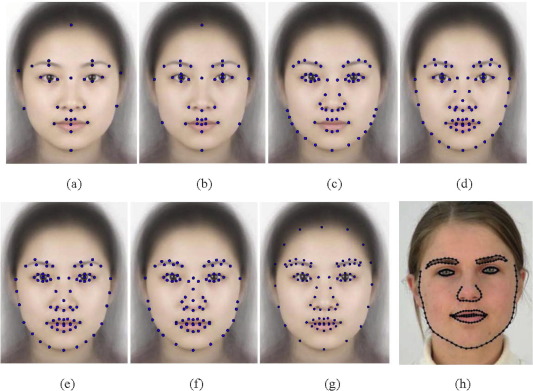
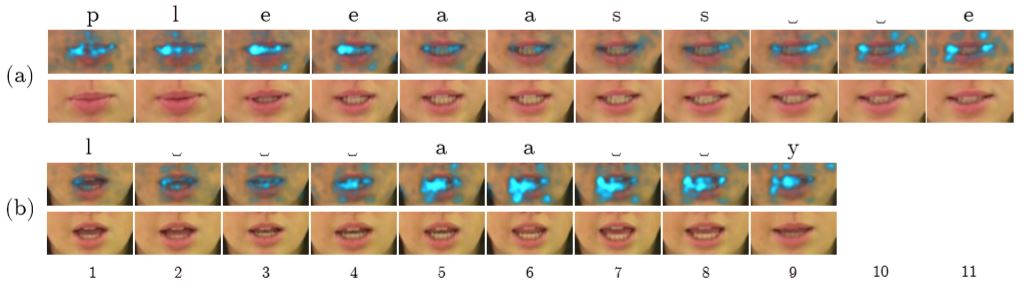


Fig. 6. A facial landmark model design by Chen et al. [14]

### Saliency and Saliency Maps

Saliency refers to how different a certain location is from its surround in color, orientation, motion , depth, etc. [40] Hence, A saliency map highlights the unique pixels of a given image and simplifies an image into something that is easier to process or analyse [53]. It is used for the purpose of selective attention in image processing [52].

LipNet applied saliency visualization techniques (Fig. 7) to interpret the system’s learned behavior, showing that the model pays attention to the movement of the lips which is phonologically significant [2].

Fig. 7. Saliency maps for please and lay, showing the places where LipNet is focusing [2].

### Data Augmentation

The more data a machine learning algorithm has access to, the more effective it can be. [60] Data augmentation is a technique used in machine learning tasks, such as image classification, to increase the training dataset size and avoid overfitting [35]. It has been proven to provide promising ways to boost the accuracy of classification tasks [60]. Some of the forms of data augmentation are cropping, resizing, horizontal and vertical flipping or mirroring, and intensity alteration [38][68].

Krizhevsky et al. (2012) incorporated horizontal reflections and image intensity alteration in their data augmentation phase. With horizontal reflections alone, the training set was increased by a factor of 2048 [41]. Zeiler et al. (2013) combined data augmentation with a stochastic pooling strategy and achieved a state-of-the-art performance of the datasets used, relative to other approaches that did not integrate data augmentation [68]. Karpathy et al. (2014), in studying the performance of CNNs in large-scale video classiﬁcation, took advantage of data augmentation to successfully reduce the effects of overfitting. The data augmentation phase involved cropping the center region of the image, resizing them, randomly sampling a region of a certain size, and finally flipping the images horizontally [38].

### Key Frame Extraction

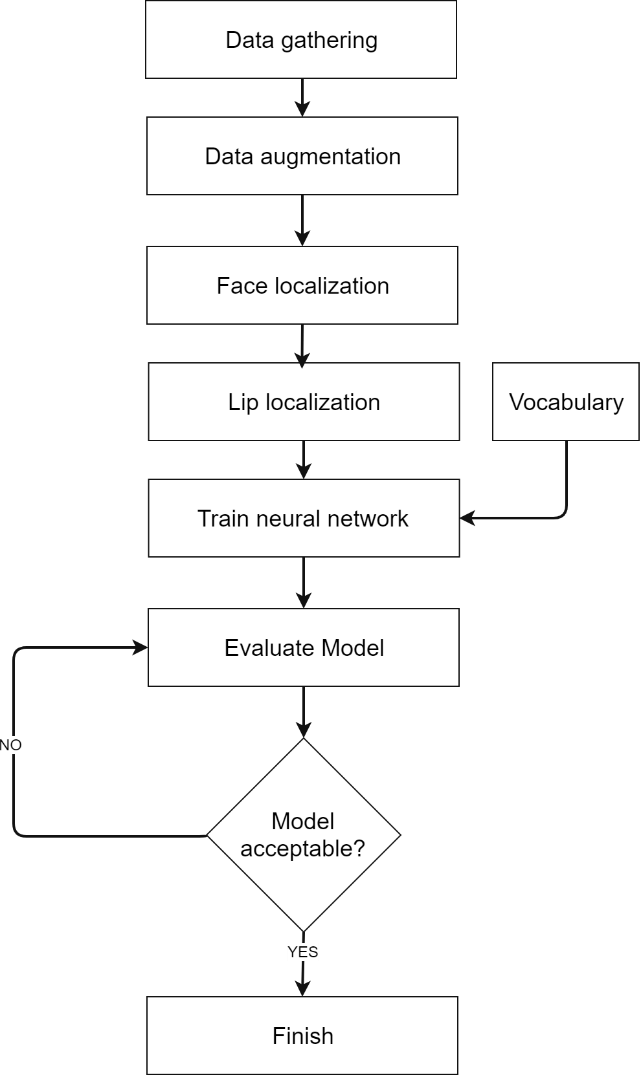
LipNet used the GRID dataset, which contain all 3 second videos with a frame rate of 25 frames-per-second. They processed the videos with DLib face detector, and the iBug facelandmark predictor with 68 landmarks. [2] Using the predicted landmarks, affine transformation was applied to extract a mouth-centered crop of size 100x50 pixels per frame. In affine transformation, all parallel lines in the original image will still be parallel in the output image [49].

Gutierrez et al. (2016) used the MIRACL-VC1 data set which contains 15 speakers, each uttering 10 words and 10 phrases for 10 times. The dataset consists of 3000 sequences of varying lengths of images with size 640x480 pixels, recorded at 15 frames-per-second. In their processing step, they utilized DLib and OpenCV and applied facial detection to crop each image to include only the face of the speaker. All the facial crops are of size 90x90 pixels in order to create a uniform input data sequence for the model [31].

In the study titled “Lipreading with Long Short-term Memory” by Wand et al. (2016), the GRID dataset is used. The frame-level alignments provided by the dataset were used to extract word-level segmentations of the video. A 40x40 pixel window of the mouth area is extracted from each video frame. The researchers used Mathematical function FindFaces[] to localize to the face area. After extraction of mouth area, the window is converted to greyscale and the contrast was maximized. All pixel values were remapped to [0,1] interval [59].

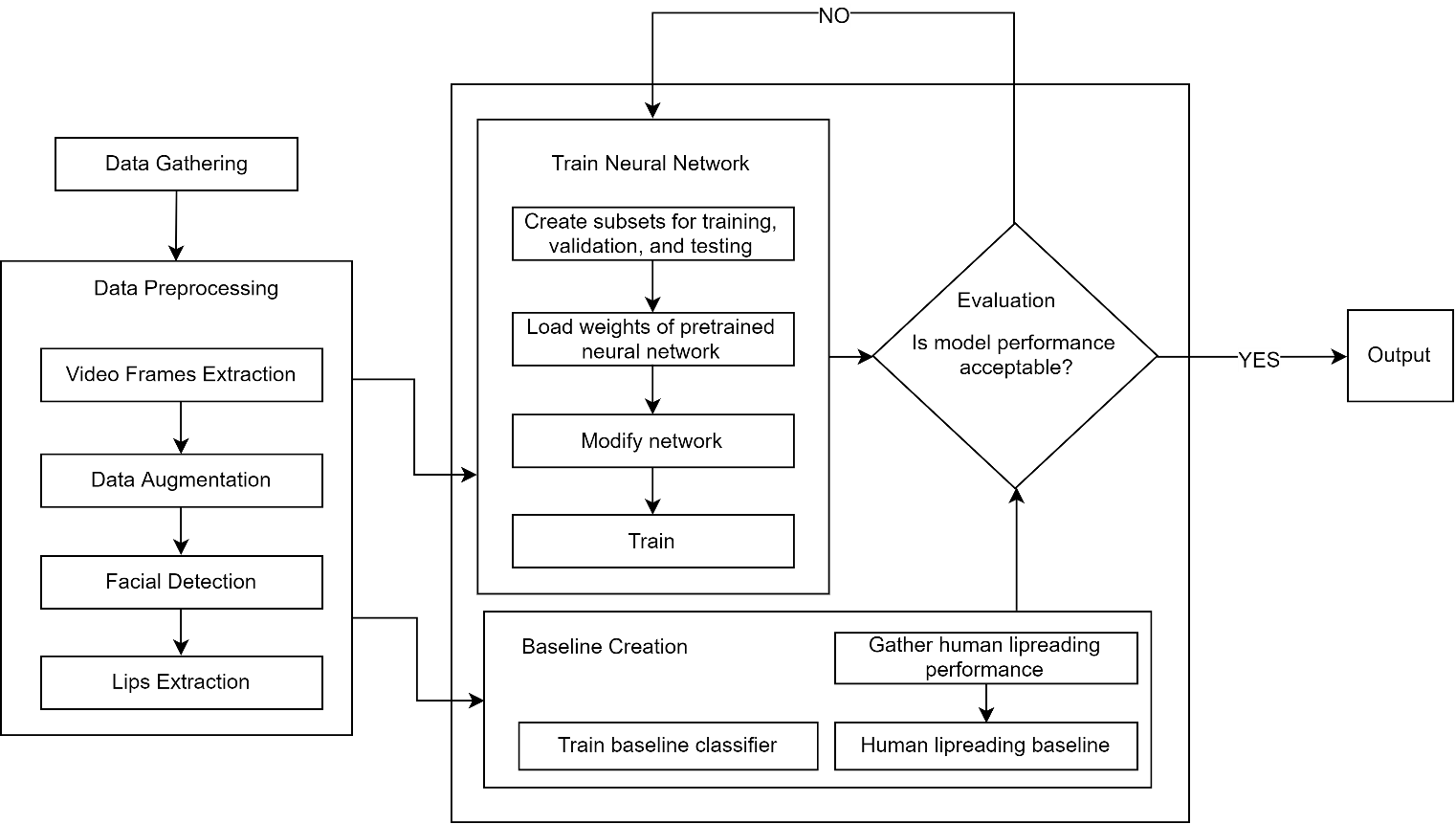
## Theoretical Framework

A framework used by Hassanat et al. (2009), typical visual speech recognition systems consists of three major processes: detecting and localizing faces, localizing the lips, followed by lipreading. The accuracy of the system depends heavily the extracted features and accurate lip localization [33]. LipNet used data augmentation with simple transformations such as regular and horizontal image mirroring [2].



# RESEARCH DESIGN AND METHODOLOGY

## Conceptual Framework



## Methodology

### Data Gathering

The data gathered contains video clips of twenty speakers uttering phrases and words [See Fig.7] ten times. The dataset contains a total of twenty-four different words. The vocabulary of words are selected according to the syllabication rules provided by the KWF (Fig. 3.), where each rule covers four words in the vocabulary. The words are explicitly chosen to maximize viseme difference. Furthermore, each instance of this dataset consists of a speaker uttering a word or a phrase once. A total of 20 × 10 × 24 = 4800 video clips will constitute the dataset for this research, before video frame extraction. The data was gathered by filming Filipino students inside the college campus of Ateneo de Davao University, in a semi-controlled environment where the distance from the camera to the speaker was maintained to be one meter.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rule # | Words | | | |
| 1 | aso | sila | almusal | basura |
| 2 | baon | paa | baunan | aasa |
| 3 | ambon | tigre | serbisyo | bulaklak |
| 4 | breyslet | transportasyon | ekskursiyon | inspirasyon |
| 5 | timpla | sobre | miyembro | balandra |
| 6 | ekstra | abstrak | eksplorasyon | transkripsiyon |

Fig.7 Words and phrases uttered by the selected speakers

### Data Preprocessing

The data collected from different college Filipino students of Ateneo de Davao University were preprocessed to reduce training time and improve the model performance. This phase is done in Python using the OpenCV library.

#### Video Frames Extraction

This phase is done for the purpose of reducing training time, for the reason that images take less time to train. A caveat of this kind of data is that each data instance contains multiple frames and each instance is of variable length depending on the speaking speed of the speaker and length of the word/phrase, which may require recurrent neural networks to boost performance.

#### Data Augmentation

This technique is done by viewing the speaker at different angles or positions in the frame to achieve a higher performance, since deep networks need large amounts of data.

Given the relatively small size of this research’s target data set, one way that can be done to combat overfitting is to augment the existing data via random transformations that still yield believable-looking images in a sequence. This exposes the model to more aspects of the data and generalize better. However, a larger dataset is still preferable because augmented data is nonetheless dependent on existing data and can remix existing information [16].

#### Facial Detection and Lips Extraction

Each image or video frame went under both processes sequentially. Facial detection is done first to ease the lips extraction process, since the face is a more visible element in the surface of the image.

### Train Neural Network

After extracting the lips, the data is then fed to a neural network for training. A neural network is used because recent researches produce higher quality outputs compared to more traditional approaches [2, 17, 18, 31]. The implementation of the network will be done with Keras with a TensorFlow backend.

#### Segregate dataset into training, validation, and testing sets

The entire dataset of video frames are segregated into training, validation, and testing sets, which are fed and tested on the neural network model.

#### Transfer Learning

Transfer learning will be used in building the model. This is done to account for limited resources. As per practice, the last few layers will be replaced in the process. The model used will be coming from ones used in ImageNet competitions. The neural network used for transfer learning will be the VGG16 model, which was developed by the Visual Geometry Group at Oxford for the ImageNet Large Scale Visual Recognition Challenge in 2014 (ILSVRC-2014 competition) [34].

#### Modify Network

The neural network’s different hyperparameters will be tuned and retuned according to the model performance. As per practice, the output layer is removed, and the network is used as a fixed feature extractor for the new data set and the first few layers are frozen. Network layers such as Convolutional layers and Long-short term memory layers will be placed on top of pretrained model.

#### Train

Multiple models with different architectures and network configurations will be trained experimentally over multiple iterations. The number of epochs, hyperparameters, and image processing techniques used in training will change depending on evaluation results of preceding models. During training, early stopping will be used to avoid overfitting the model, stopping the training process when there are no more improvement gains from a number of epochs. Model checkpointing will also be used in order to save the best model during training. Both early stopping and model checkpointing are available in Keras.

### Baseline Creation

Baselines will serve as a way to gauge the main model’s performance. The proponents will use three baselines for evaluation and comparison: random baseline, human lipreading performance, baseline model.

#### Random Baseline

Since the dataset used in this research is balanced (24 classes with equal number of data instances), a simple random baseline can be used for model performance evaluation as this yields the same results with a ZeroR baseline, which predicts the majority class.

#### Gather Human Lipreading Performance

Another set of students from the Ateneo de Davao University college campus were asked to lipread the speakers from the data collected. The students will lipread from a sequence of video frames that will be fed into the model rather than from videos to ensure a more accurate measure of performance when comparing the results with the model.

#### Human Lipreading Baseline

The data gathered from the human lipreading by students of Ateneo de Davao University is assessed to create a baseline for this study.

#### Train Baseline Classifier

The baseline model used in this research will be a product of one of the more relatively successful iterations, decided upon by several evaluation metrics to be mentioned below. The baseline will have a relatively simple architecture. The baseline used by Gutierrez et al. (2016) is a neural network consisting of a Convolutional Neural Network and a Long Short-Term Memory Recurrent Neural Network (CNN + LSTM), which produces a single output, making it a many-to-one.

### Evaluation

The models will primary be evaluated based on their training and validation loss and accuracy over a number of epochs. The model’s recall and precision will also be taken into account. Visualization of classification errors and performances will be done through the generation of confusion matrices and graphs.

REFERENCES

[1] Allophone | linguistics | Britannica.com: .

[2] Assael, Y.M., Shillingford, B., Whiteson, S. and de Freitas, N. 2016. LipNet: End-to-End Sentence-level Lipreading. (2016), 1–13.

[3] Bautista, L. and Bolton, K. 2008. *Philippine English: Linguistic and Literary*. Hong Kong University Press.

[4] Bear, H.L. 2017. Decoding visemes: improving machine lipreading (PhD thesis). July (2017).

[5] Bear, H.L. and Bear, H.L. 2017. Understanding the visual speech signal. June (2017), 12–15. DOI:https://doi.org/10.13140/RG.2.2.22035.48165.

[6] Bear, H.L., Cox, S.J. and Harvey, R.W. 2017. Speaker-independent machine lip-reading with speaker-dependent viseme classifiers. August (2017).

[7] Bear, H.L., Cox, S.J. and Harvey, R.W. 2017. Speaker-independent machine lip-reading with speaker-dependent viseme classifiers. September (2017).

[8] Bear, H.L., Harvey, R.W., Theobald, B.-J. and Lan, Y. 2017. Which phoneme-to-viseme maps best improve visual-only computer lip-reading? (2017). DOI:https://doi.org/10.1007/978-3-319-14364-4\_22.

[9] Cai, N., Lin, Z., Zhang, F., Cen, G. and Wang, H. Detecting facial landmarks in the video based on a hybrid framework. 1–8.

[10] Caplier, A. 2001. Lip detection and tracking. *Proceedings - 11th International Conference on Image Analysis and Processing, ICIAP 2001* (2001), 8–13.

[11] Carrera, P.F. de 2010. Face Recognition Algorithms. (2010), 1–66.

[12] Cascade Classification — OpenCV 2: 2015. *https://docs.opencv.org/2.4/modules/objdetect/doc/cascade\_classification.html*. Accessed: 2018-07-03.

[13] Chang-yeon, J. 2008. Face Detection using LBP features. (2008), 1–4.

[14] Chen, F., Xu, Y., Zhang, D. and Chen, K. 2015. 2D facial landmark model design by combining key points and inserted points. *Expert Systems with Applications*.

[15] Chen, T., Xu, R., He, Y. and Wang, X. 2017. Deep Learning in Neural Networks: An Overview. *Expert Systems with Applications*. 72, (2017), 221–230. DOI:https://doi.org/10.1016/j.eswa.2016.10.065.

[16] Chollet, F. *Deep Learning with Python*.

[17] Chung, J.S., Senior, A., Vinyals, O. and Zisserman, A. 2016. Lip Reading Sentences in the Wild. (2016). DOI:https://doi.org/10.1109/CVPR.2017.367.

[18] Chung, J.S. and Zisserman, A. Out of time: automated lip sync in the wild. i.

[19] Cootes, T. and Edwards, G. 1998. A comparative evaluation of active appearance model algorithms. *British Machine Vision Conference*. (1998).

[20] Eveno, N., Caplier, A. and Coulon, P.Y. 2004. Accurate and quasi-automatic lip tracking. *IEEE Transactions on Circuits and Systems for Video Technology*. 14, 5 (2004), 706–715. DOI:https://doi.org/10.1109/TCSVT.2004.826754.

[21] Eveno, N., Caplier, a. and Coulon, P.-Y. 2002. A parametric model for realistic lip segmentation. *7th International Conference on Control, Automation, Robotics and Vision, 2002. ICARCV 2002.* 3, (2002), 1426–1431. DOI:https://doi.org/10.1109/ICARCV.2002.1234982.

[22] Face Detection using OpenCV and Python: A Beginner’s Guide: 2017. *https://www.superdatascience.com/opencv-face-detection/*. Accessed: 2018-07-03.

[23] Facial Landmark Detection \_ Learn OpenCV: 2015. *https://www.learnopencv.com/facial-landmark-detection/*. Accessed: 2018-01-06.

[24] Fisher, C.G. 1968. Confusions Among Visually Perceived Consonants. *Journal of Speech Language and Hearing Research*. 11, 4 (Dec. 1968), 796. DOI:https://doi.org/10.1044/jshr.1104.796.

[25] Froba, B. and Kulbekc, C. 2001. Real-Time Face Detection Using Edge-Orientation Matching \_ SpringerLink. *Real-Time Face Detection Using Edge-Orientation Matching* (Berlin, Heidelberg, 2001).

[26] Fu, Y., Zhou, X., Liu, M., Hasegawa-Johnson, M. and Huang, T. 2006. Lipreading by locality discriminant graph — University of Illinois at Urbana-Champaign.

[27] Garg, A. and Noyola, J. Lip reading using CNN and LSTM.

[28] Gomez, E., Travieso, C., Briceno, J. and Ferrer, M.A. 2002. Biometric identification system by lip shape. *36th Annual International Carnahan Conference on Security Technology* (2002), 39–42.

[29] Graves, A., Fernandez, S., Gomez, F. and Schmidhuber, J. 2006. Connectionist Temporal Classification : Labelling Unsegmented Sequence Data with Recurrent Neural Networks. *Proceedings of the 23rd international conference on Machine Learning*. (2006), 369–376. DOI:https://doi.org/10.1145/1143844.1143891.

[30] Guevara, R.C.L., Ph, D., Garcia, I.D.S., Santos, T.I.D., Ma, R. and Ph, D.N. 1967. A Computational Approach to Filipino Speech Rhythm. (1967).

[31] Gutierrez, A. and Robert, Z.-A. Lip Reading Word Classification.

[32] Hassanat, A. 2010. Color-based Lip Localization Method Color-based Lip Localization Method. April (2010). DOI:https://doi.org/10.1117/12.850629.

[33] Hassanat, A.B.A. 2009. Visual Speech Recognition. (2009).

[34] ImageNet Classification: .

[35] Inoue, H. 2017. D ATA A UGMENTATION BY P AIRING S AMPLES. (2017), 1–10.

[36] Jesorsky, O., Kirchberg, K.J. and Frischholz, R.W. 2001. Robust Face Detection Using the Hausdorff Distance. *In Proc. Third International Conference on Audio- and Video-based Biometric Person Authentication*. June (2001), 90–95. DOI:https://doi.org/10.1007/3-540-45344-X\_14.

[37] Kadlec, J. 2004. *Student EEICT 2004 : proceedings of the 10th conference and competition : papers written by postgraduate students*. Zdeněk Novotný.

[38] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L. 2014. Large-scale video classification with convolutional neural networks. *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. (2014), 1725–1732. DOI:https://doi.org/10.1109/CVPR.2014.223.

[39] Kass, M., Witkin, A. and Terzopoulos, D. 1988. Snakes: Active contour models. *International Journal of Computer Vision*. 1, 4 (1988), 321–331. DOI:https://doi.org/10.1007/BF00133570.

[40] Koch, C. and Ullman, S. 1987. Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. *Matters of Intelligence*. (1987), 115–141. DOI:https://doi.org/10.1007/978-94-009-3833-5\_5.

[41] Krizhevsky, A., Sutskever, I. and Hinton, G.E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. (2012).

[42] Lalitha, S.D. 2016. A Study on Lip Localization Techniques used for Lip reading from a Video. 11, 1 (2016), 611–615.

[43] Mahdi, W., Werda, S. and Ben Hamadou, A. 2008. A hybrid approach for automatic lip localization and viseme classification to enhance visual speech recognition. *INTEGRATED COMPUTER-AIDED ENGINEERING*. 15, 3 (2008), 253–266.

[44] Milborrow, S. and Nicolls, F. 2008. Locating facial features with an extended active shape model. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2008), 504–513.

[45] Muller, N., Magaia, L. and Herbst, B.M. 2004. Singular Value Decomposition, Eigenfaces, and 3D Reconstructions. *SIAM Review*. 46, 3 (2004), 518–545. DOI:https://doi.org/10.1137/S0036144501387517.

[46] Nespor, M. and Mehler, J. 1989. timed Languages. 1 (1989), 1147–1159.

[47] Noche, P. 2016. Syllabication of Filipino Words. (2016), 1–7.

[48] Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H.G. and Ogata, T. 2014. Lipreading using Convolutional Neural Network. September (2014), 1149–1153.

[49] OpenCV\_ Affine Transformations: 2015. *https://docs.opencv.org/3.1.0/d4/d61/tutorial\_warp\_affine.html*.

[50] Parmar, D.N. and Mehta, B.B. 2013. Face Recognition Methods & Applications. *Int. J. Computer Technology & Applications*. 4, 1 (2013), 84–86.

[51] Rekik, A. 2014. A New Visual Speech Recognition Approach for RGB-D Cameras. October (2014). DOI:https://doi.org/10.1007/978-3-319-11755-3.

[52] Saliency map - Scholarpedia: 2017. *http://www.scholarpedia.org/article/Saliency\_map*. Accessed: 2018-01-07.

[53] Simonyan, K., Vedaldi, A. and Zisserman, A. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. (2013), 1–8.

[54] Stress-timed and syllable-timed languages | InTuition Languages: 2017. .

[55] The difference between speaker-dependent and speaker-independent recognition software - SpeechAngel: 2016. *https://speechangel.com/2016/05/04/difference-speaker-dependent-speaker-independent-recognition-software/*. Accessed: 2018-01-05.

[56] The Sound of English: Glottal Stop: .

[57] Wand, M. and Koutn, J. Lipreading With Long Short-term Memory.

[58] Wand, M., Koutník, J. and Schmidhuber, J. 2016. Lipreading with Long Short-Term Memory. *PhD Proposal*. 1, 5 (2016), 248–256. DOI:https://doi.org/10.1088/1751-8113/44/8/085201.

[59] Wand, M., Koutník, J. and Schmidhuber, J. 2016. Lipreading with Long Short-Term Memory. (2016).

[60] Wang, J. and Perez, L. 2017. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. *Unpublished*. (2017).

[61] Wang, Y.-Q. 2014. An Analysis of the Viola-Jones Face Detection Algorithm. *Image Processing On Line*. 4, (2014), 128–148. DOI:https://doi.org/10.5201/ipol.2014.104.

[62] Werda, S., Mahdi, W. and Hamadou, A. Ben 2007. Lip localization and viseme classification for visual speech recognition. *Internation Journal of Computing & Information Sciences*. 5, 1 (2007), 62–75.

[63] Van Wettum, Y.C. 2016. Facial Landmark Tracking on a Mobile Device. December (2016), 1–16.

[64] Wilson, P.I. and Fernandez, J. 2006. Facial Feature Detection Using Haar Classifiers. *Journal of Computing Sciences in Colleges*. 21, 4 (2006), 127–133. DOI:https://doi.org/10.1109/CVPR.2001.990517.

[65] Wu, Y.W.Y., Liu, H.L.H. and Zha, H.Z.H. 2004. A new method of detecting human eyelids based on deformable templates. *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*. 1, (2004), 604–609. DOI:https://doi.org/10.1109/ICSMC.2004.1398366.

[66] Yargic, A. and Dogan, M. 2013. A lip reading application on MS Kinect camera. *2013 IEEE International Symposium on Innovations in Intelligent Systems and Applications, IEEE INISTA 2013* (2013).

[67] Yuille, A.L., Hallinan, P.W. and Cohen, D.S. 1992. Feature extraction from faces using deformable templates. *International Journal of Computer Vision*. 8, 2 (1992), 99–111. DOI:https://doi.org/10.1007/BF00127169.

[68] Zeiler, M.D. and Fergus, R. 2013. Stochastic Pooling for Regularization of Deep Convolutional Neural Networks. (2013), 1–9.