Recurrent Neural Networks for Decoding Lip Read Speech

Souheil Fenghour
Department of Engineering
London South Bank University
London, United Kingdom
fenghous@lsbu.ac.uk

Daqing Chen
Department of Engineering
London South Bank University
London, United Kingdom
chend@lsbu.ac.uk

Perry Xiao
Department of Engineering
London South Bank University
London, United Kingdom
perry.xiao@lsbu.ac.uk

ABSTRACT

The success of automated lip reading has been constrained by the inability to distinguish between homophemes, which is when words have different characters producing the same lip movements (e.g. ”time” and ”some”) [1], despite being intrinsically different. One word can often have different phonemes (units of sound) [2] [3] producing exactly the viseme or visual equivalent of phoneme for a unit of sound. Through the use of a Long-Short Term Memory Network with word embeddings, we can distinguish between homopheme words or words that produce identical lip movements. The neural network architecture achieves a character accuracy rate of 77.1% and a word accuracy rate of 72.2%.

Keywords

Artificial Intelligence; Deep Learning; Lip Reading; Speech Recognition; Natural Language Processing; Recurrent Neural Networks; Long-Short Term Memory Networks; Word Embeddings

1. INTRODUCTION

Visual Speech Recognition or Lip Reading plays an important role in human communication - especially in noisy environments where audio speech recognition may be difficult. It is can also be extremely useful for people who's hearing is impaired, for those who are autistic and for those suffering from language impairment not to mention that is would serve as a useful too in assisting the police decipher CCTV footage of people speaking when audio is unavailable [4] [5] [6] [7] [8].

Automated Lip Reading remains a very challenging task and one that is made more challenging not only when there is no audio available for assistance. Numerous attempts of recent have been made to automate lip reading through a variety of methodologies including Hidden Markov Models [9], Support Vector Machines [10] and Neural Networks [11] [12] [13] [14].

Automated Lip Reading has encountered many obstacles such as the insufficient supply of datasets that would be needed to train effective models, the presence of facial features and poor lighting that can inhibit feature extraction in automated lip reading systems as well as the inability to distinguish between homopheme words or words that produce identical lip movements despite being different and sounding different and this particular problem is the focus of the paper.

Automated lip reading normally requires algorithms for image processing to map lip movements to speech articulators and language algorithms to further predict what has been said given the background context. The issue of feature extraction is not discussed in this paper as we are purely focused on the issue of deciphering what has been said given the precise lip movements produced by the speaker and how one can figure out the word's spoken given that one set of lip movements could correspond to several words.

2. RELATED WORKS

A variety of non-deep learning based methodologies have been used to automate lip reading including Hidden Markov Models and Support Vector Machines and such methods make up the vast majority of approaches for automated lip reading. They are however far too extensive to review in this paper, but interested readers can read Zhou et al's [15] work for an extensive review of such methods. Deep Learning approaches to visual speech recognition have been focused on word classification. Approaches include Wand et al [14], Garg et al [13], Chung and Zisserman [12] and LipNet [11].

Wand et al. in 2016 [14] used Long Short Term Memory (LSTM) networks for lip reading, achieved an accuracy rate of 79.6% for word classification, tough the one major limitation to their approaching that it was speaker dependant and would not achieve as good an accuracy rate when evaluated on other speakers.

Garg et al. (2006) [13] used a pre-trained Convolutional Neural Network (CNN) based system to recognise words and phrases from the MIRACL-VC1 dataset which consists of just 10 words and phrases. An LSTM is also trained though the CNN and LSTM need to be trained separately. The overall system managed a limited accuracy and the dataset used in their approach is relatively small and insufficient to train a model that could cover a wide enough range of subjects.

Later neural network based approaches for lip reading have deployed deep stacked networks consisting neural networks in a stacked configuration where they can be trained

WOODSTOCK '97 El Paso, Texas USA

DOI: 10.475/123_4

simultaneously. Chung and Zisserman(2017) [12] and LipNet [11] have made attempts to lip read entire sentences with both methods achieving good accuracy rates on their own respective datasets. Despite recording good accuracy both approaches are limited in their ability to classify visemes correctly as both systems are word-based classification model not trained specifically for the tasks of phoneme or viseme classification furthermore, both approaches are still limited in their ability to distinguish between words with identical visemes i.e. homopheme words.

The neural network that is the subject of this paper is an LSTM based network designed to predict what word may be present given the combination of visemes that are uttered by a speaker and is this all performed through the use of word embeddings which allow us to predict the word spoken by context recognition. In a real-life situation, one would still need to need to be able to decode which visemes have been uttered by the speaker given their lip movements but this is not the main focus of this paper.

The neural network structure is modelled according to neural machine translation [16] where stacked Recurrent Neural Networks (RNNs) are used to convert sequences of text from one language into another. Neural machine translation systems follow an encoder-decoder structure whereby the encoder reads an input sentence to encode it into a fixed-length vector, and the decoder would then output a translation from the vector having been trained to maximise the probability of the ”correct translation” given an input sentence. One of the main advantages of the encoder-decoder models is its ability to deal with varying lengths of input and output text sequences [17].

One significant difference between machine translation and homopheme classification is that the former tends to be one-to-one mapping whereas the latter requires one-to-many mapping because one combination of visemes can be mapped to many different words but like machine translation, context is required to decipher the identity of a spoken word through conditional probability.

3. METHODOLOGY

In this Section, we explain the fundamental units of speech namely phonemes and visemes, and how they can be used to classify what is spoken upon their recognition. Additionally the dataset used to train and test our own architecture is explained as are the algorithms used for evaluating the accuracy if the architecture.

3.1 Phonemes and Visemes

A viseme is the most fundamental unit of visual speech and the visual equivalent of a phoneme, the latter of which is a spoken unit of speech that can be represented by an acoustic signal. According to Hazen [18], there are roughly 40 phonemes in the English language with only around a dozen distinguishable visemes, meaning that several sounds can produce identical lip movements and thus resulting in words that look the same when spoken, i.e. homophemes or homovisemes which happen to be a more common occurrence than homophone words, i.e. those that sound the same when spoken.

There are a variety of conventions which have been used to classify phonemes and visemes when analysing visual speech including and they all differ in their definitions of how many precise phonemes or visemes there are. In this paper, we will be using the viseme convention outlined in Table 1 which is that of Lee [19] has been shown visually in Figure 1 and appears to be the most favoured for visual speech recognition and for phonemes, we have used the convention Carnegie Mellon University Pronouncing Dictionary [20].

It was Alexander Graham Bell who first hypothesized that multiple phonemes may be visually identical on a given speaker. Because one viseme can generate multiple phonemes, the mapping of visemes to phonemes represents a one-to-many relationship [2] [21].

The fact that multiple phonemes may share the same viseme poses problems for lip reader interpreters because you can have so many words producing the exact same lip movements and would therefore look the same so lip readers have to decipher exactly which word was spoken when there is no audio present. This is one of the many challenges faced in automated lip reading.

Words consist of phonetic symbols or phonemes which can in turn be mapped to visemes. For our neural network architecture words will be represented as combinations of visemes and every distinct combination of visemes will have its own distinct lip movements. A full visual speech recognition system would consist of an initial feature recognition stage for decoding the words that have been uttered given the representation of lip movements but this paper is focused on decoding the presence of a word or sequence of words upon the detection of distinct visemes which are based on lip movements.

Table 1: Lee and Yook's viseme convention with vowels and consonants [19]

Viseme Class	Viseme Type	Phonemes Set
p	consonant	b, p, m
t	consonant	d, t, s, z, th, dh
k	consonant	g, k, n, ng, l, y, hh
ch	consonant	jh, ch, sh, zh
f	consonant	f, v
w	consonant	r, w
iy	vowel	iy, ih
ey	vowel	eh, ey, ae
aa	vowel	aa, aw, ay, ah
ah	vowel	ah
ao	vowel	ao, oy, ow
uh	vowel	uh, uw
er	vowel	er
s	silent character	sil

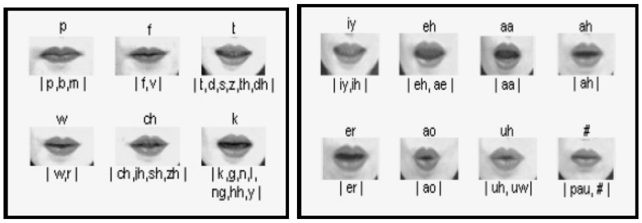

Figure 1: The six consonant visemes on the left and the 7 vowel visemes and silent viseme on the right. [19]

3.2 Dataset

The TIMIT corpus is an audio-visual dataset consisting of 630 speakers each speaking 10 different sentences giving a total of 6300 sentences available for training and testing. The speakers utter vocabulary covering the eight major dialects of American English and the overall speech has a balanced distribution of phonemes [22].

The data within the corpus consists of two parts; the first being videos that are approximately 30 seconds long that have been converted into image frames having been sampled at 25 frames per second and the second being the subtitles consisting of words that are spoken at each time step. The subtitles are word transcriptions that will have been subsampled at 16 kHz and there will be subtitles listing each word spoken, the starting time for that word and the stopping time for that word.

For the purposes of performing homopheme detection using RNNs, the first part of the corpus will not be necessary. It is assumed that words with identical sequences of visemes share the exact same lip movements which is a theory put forward by Alexander Graham Bell [2] [21]. It is the sequence of words that will be required for performing the simulations where each word will be treated like a label and its combination of visemes will be treated as a class.

Because there are repeated sentences with one or more speakers uttering the same sentence as another, the overall dataset actually consists of 2363 distinct sentences with a vocabulary list of 6099 different words of which there are 4764 distinct viseme combinations.

3.3 Accuracy Metrics

The two of the metrics evaluating the accuracy of our architecture are character error rate(CER) and word error rate(WER) - both commonly used metrics for evaluating the accuracy of speech recognition systems.

In determining misclassifications, one has to compare the decoded speech to the actual speech and the alterations that are required to get from the decoded sentence to the actual sentence. If we look at Eq.1, N is the total number of words in the actual speech, S is the number of substitutions made for wrong classifications, I represents the number insertions made for words not picked up while D is the number of deletions being made for decoded words that should not be present. The word error rate WER is defined as the ratio of incorrect words decoded to the total number of words in a sample(given by Eq. 1).

Character error rate CER is calculated the same way as WER except that characters are evaluated instead of words. Furthermore, the word accuracy rate WAR and character accuracy CAR can be calculated by subtracting the either error rate from the number 1 respectively according to Eq. 2. Tables 3 and 4 give examples of how the character and word accuracies can be calculated.

$$WER = \frac{(S+D+I)}{N} \tag{1}$$

$$WAR = 1 - WER \tag{2}$$

Table 2: Character error rates calculations for different phrases.

Case 1	Case 2	S	D	I	N	CAR(%)
bin blue in o six now	bin blue at l six now	3	0	0	3	85.8
bin blue a x e again	bin blue at s three again	1	0	5	6	76.0
lay white at e zero please	lay red in e zero please	5	2	0	7	70.8

Table 3: Word error rates calculations for different phrases.

Case 1	Case 2	S	D	I	N	WAR(%)
bin blue in o six now	bin blue at l six now	2	0	0	6	66.7
bin blue a x e again	bin blue at s three again	3	0	0	6	50.0
lay white at e zero please	lay red in e zero please	2	0	0	6	66.7

3.4 Neural Network Architecture

The neural network architecture (Figure 2) we are using is modelled according to neural machine translation and it is there for a stacked LSTM with word embeddings, a repeat vector and time-distributed network following the "encoder-decoder" model (Figure 3).

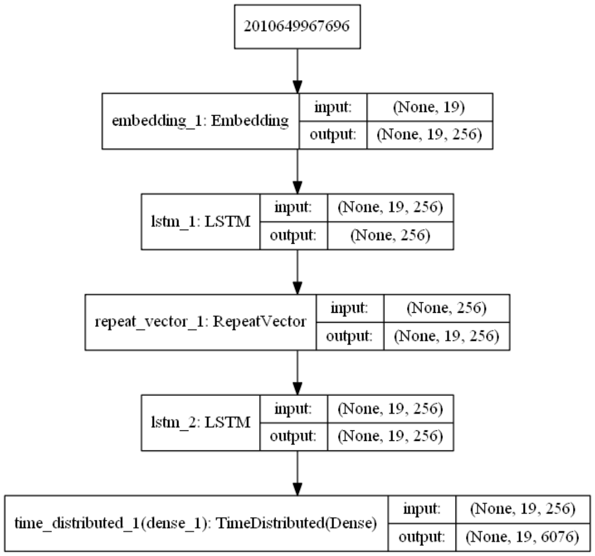

Figure 2: Structure and dimensions of the stacked LSTM configuration with the word embeddings, repeat vector and time distributed network.

Out of the 6300 sentences available to us, 90% of them will be used for training while the remainder will be used for testing. All 6300 sentences will have been converted to sequences of viseme combinations beforehand and each combination of visemes will be assigned a class label. subsequently, every sentence would be treated as a sequence of classes (Table 4 shows the split between testing and training).

A sequence of viseme combinations forms the input of the architecture while the output is a sequence of words to be

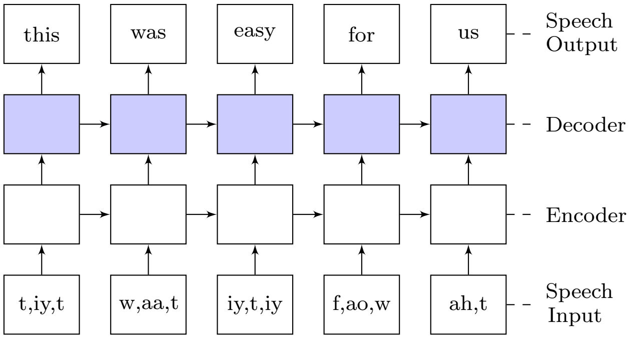

Figure 3: Diagram of stacked RNN structure showing the inputs with visemes and outputs with decoded words.

Table 4: Train/Test Data for homopheme classification.

Situation	Number of videos used
Training	5760
Testing	630

predicted by the network. In the same way that every viseme combination will be treated like a class, every possible word that could be predicted by the network will also be treated like a class.

For an encoder-decoder framework, an input sentence of the form of a sequence of word vectors $\{x_1, \ldots, x_t\}$ where x_t corresponds to a vector, into a vector c with hidden state h_t at time t. The vector c is generated from the sequence of hidden states while f and q are non-linear variables.

$$h_t = f(x_t, h_{t-1}) \tag{3}$$

$$c = q(\{h_1, \ldots, h_t\}) \tag{4}$$

The decoder is trained to predict next word y_t given the context vector c and all the previously predicted words $\{y_1, \ldots, y_t\}$ The decoder defines a probability $p(y)$ over the translation y by considering the joint conditional probability of all other previous words.

$$p(y) = \prod_{t=1}^{T} a_i p(y_t | \{y_1, \ldots, y_t\}, c) \tag{5}$$

$$p(y_t | \{y_1, \ldots, y_t\}, c) = g(y_{t-1}, s_t, c) \tag{6}$$

4. RESULTS

The overall stacked neural network architecture was trained on all 5670 sentences from the training set consisting of viseme combinations that were labelled by sentences composed of actual words. The network predicted the words present when combinations of visemes were inputed. Table 5 shows the overall results achieved once the network had gone through 400 epochs of iterations with sentences being grouped into batches of 60 for each iteration where an average WER and CAR of 72.2% and 77.1% respectively were achieved.

If we analyse a sample of the results in Table 6, we can see that some of the sentences follow unusual sequences where there are for example repeated words such as ”that that” or ”it it” and there are words decoded by the network that do not correspond to the actual words in the input sequence so there is a need to further improve the overall accuracy of the architecture.

Table 5: Results for average word-error rates and character-error for word classifications in sentences evaluated by our architecture.

	Epochs	Sentences	WAR(%)	CAR(%)
Architecture	400	630	72.2	77.1

Table 6: A sample of how some of the sentences were decoded by the neural network during the testing phase.

Decoded Phrase	Actual Subtitle	CAR(%)	WAR(%)
academic aptitude guarantees your diploma	academic aptitude guarantees your diploma	100.0	100.0
the misprint provoked an immediate disclaimer	the misprint provoked an immediate disclaimer	100.0	100.0
do atypical farmers grow oats	do atypical farmers grow oats	100.0	100.0
the surplus shoes were sold at a discount price	the surplus shoes were sold at a discount price	100.0	100.0
a tube a a a a the of the of	quite often honeybees form a majority on the willow catkins	28.8	10.0
that that it it shrinking shrinking faster	but that explanation is only partly true	20.0	14.3

5. CONCLUSION

We have addressed one of the major challenges faced in machine-based lip reading which is the issue of distinguishing between homopheme words or words that produce identical lip movements. It has been demonstrated that through the use of a stacked configuration of recurrent neural networks that has been tested on a dataset designed for audio-visual speech recognition, we can detect the identify of a word in an uttered sentences provided that the visemes combinations of spoken words have been accurately recognised.

Further work is required to improve the accuracy of our system and simulation results have shown that words decoded incorrectly do not share the same visemes as the true spoken words and some decoded sentences consisted of repeated words and this something that could tackled algorithmically. The efficiency of the overall architecture is an area that could be reviewed for example an ”encoder-decoder” may not be necessary given that the number of input viseme combinations matches the the number of words in each sentences meaning that we are not dealing with length variability.

6. ADDITIONAL AUTHORS

7. REFERENCES

[1] A. J. Goldschen, O. N. Garcia and E. D. Petajan. (1997). Continuous automatic speech recognition by lipreading. InMotion-Based recognition.

[2] C. G. Fisher. (1968). Confusions among visually perceived consonants. Journal of Speech, Language, and Hearing Research.

[3] F. Woodward and C. G. Barber. (1960). Phoneme perception in lipreading. Journal of Speech, Language, and Hearing Research.

[4] E. T. Auer and L. E. Bernstein. (2007). Enhanced Visual Speech Perception in Individuals with Early-Onset Hearing Impairment. Journal of Speech, Language, and Hearing Research.

[5] R. Campbell and T. E. Mohammed. (2010). Speechreading for information gathering: a survey of scientific sources. Deafness Cognition and Language Research Centre.

[6] R. Bowden et al. (2013). Recent developments in automated lip-reading. Proceedings of SPIE - The International Society for Optical Engineering.

[7] M. Bohning *et al.* (2002). Audiovisual speech perception in Williams syndrome. Neuropsychologia.

[8] J. Leybaert *et al.* (2014). Atypical audio-visual speech perception and McGurk effects in children with specific language impairment. Front Psychol.

[9] C. Neti *et al.* (2000). Audio visual speech recognition. Technical report IDIAP.

[10] K. Saenko, K. Livescu, J. Glass, and T. Darrell. (2005). Production Domain Modeling Of Pronunciation For Visual Speech Recognition. ICASSP.

[11] Y. M. Assael, B. Shillingford, S. Whiteson and N. de Freitas. (2016). LipNet: End-to-Endsentence-Level Lipreading. ICLR Conference.

[12] J. S. Chung, A. Zisserman, A. Senior and O. Vinyals. (2016). Lip Reading Sentences in the Wild. IEEE Conference on Computer Vision and Pattern Recognition.

[13] A. Garg, J. Noyola, and S. Bagadia. (2016). Lip reading using CNN and LSTM. Technical report Stanford University - CS231n project report.

[14] M. Wand, J. Koutnik, and J. Schmidhuber. (2016). Lipreading with long short-term memory. In IEEE International Conference on Acoustics, Speech and Signal Processing pp 6115-6119.

[15] Z. Zhou, G. Zhao, X. Hong and M. PietikÃd'inen. (2014). A review of recent advances in visual speech decoding. Image and vision computing.

[16] B. Bahdanau, K. Cho, and Y Bengio. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. ArXiv. 1409.

[17] Y. Wu et al. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.

[18] T. J. Hazen, K. Saenko, C. La and J. R. Glass. (2004). A segment-based audio-visual speech recognizer: data collection, development, and initial experiments. Proceedings of the 6th International Conference on Multimodal Interfaces.

[19] S. Lee and D. Yook. (2002). Audio-to-Visual Conversion Using Hidden Markov Models. In Proceedings of the 7th Pacific Rim International Conference on Artificial Intelligence: Trends in Artificial Intelligence.

[20] R. Treiman, B. Kessler and S. Bick. (2001). Context sensitivity in the spelling of English vowels. Journal of Memory and Language.

[21] F. DeLand. (1931). The story of lip-reading, its genesis and development.

[22] L. Lamel, R, H. Kassel and S. Seneff. (1989). Speech database development: Design and analysis of the acoustic-phonetic corpus. Proceedings of the DARPA Speech Recognition Workshop.