Available online at www.sciencedirect.com
ScienceDirect
Computer Speech & Language xxx (2018) xxx-xxx
www.elsevier.com/locate/csl

Comparing heterogeneous visual gestures for measuring the diversity of visual speech signals☆

Q2 Helen L. Bear[a,*], Richard Harvey[b]

[a] CVSSP, Department of Electrical Engineering, University of Surrey, Guildford GU2 7JP, UK
[b] School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK

Received 20 March 2018; received in revised form 13 April 2018; accepted 4 May 2018
Available online xxx

Abstract

Visual lip gestures observed whilst lipreading have a few working definitions, the most common two are: 'the visual equivalent of a phoneme' and 'phonemes which are indistinguishable on the lips'. To date there is no formal definition, in part because to date we have not established a two-way relationship or mapping between visemes and phonemes. Some evidence suggests that visual speech is highly dependent upon the speaker. So here, we use a phoneme-clustering method to form new phoneme-to-viseme maps for both individual and multiple speakers. We test these phoneme to viseme maps to examine how similarly speakers talk visually and we use signed rank tests to measure the distance between individuals. We conclude that broadly speaking, speakers have the same repertoire of mouth gestures, where they differ is in the use of the gestures.

Keywords: Visual speech; Lipreading; Recognition; Audio-visual; Speech; Classification; Viseme; Phoneme; Speaker identity

1. Introduction

Computer lipreading is machine speech recognition from the interpretation of lip motion without auditory support (Stafylakis and Tzimiropoulos, 2017a; Bear and Taylor, 2017). There are many motivators for wanting a lipreading machine, for example places where audio is severely hampered by noise such as an airplane cockpit, or where placing a microphone close to a source is impossible such as a busy airport or transport hub (Neti et al., 2000; Morade and Patnaik, 2014; Bear et al., 2014c)

Conventionally, machine lipreading has been implemented on two-dimensional videos filmed in laboratory conditions (Harte and Gillen, 2015; Cooke et al., 2006). More recently, such datasets have been growing in size to enable deep learning methods to be applied in lipreading systems (Chung and Zisserman, 2016; Thangthai and Harvey, 2017). Separately there has also been some preliminary work to use depth cameras to capture pose/lip protrusion information (Heidenreich and Spratling, 2016; Watanabe et al., 2016) or in the RGB colour space for more discriminative appearance features (Rekik et al., 2016). The challenge with these works are that the results achieved are yet

☆ This paper has been recommended for acceptance by Roger Moore.
* Corresponding author.
E-mail address: dr.bear@tum.de (H.L. Bear), r.w.harvey@uea.ac.uk (R. Harvey).

https://doi.org/10.1016/j.csl.2018.05.001
0885-2308/

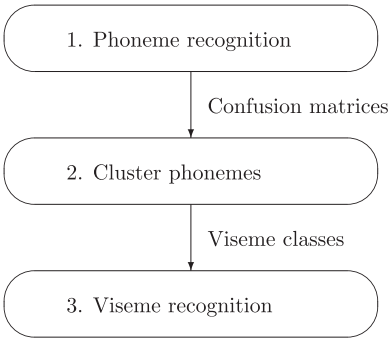

Fig. 1. Three step process for recognition from visemes. This figure summarises the process undertaken by Bear et al. (2014b).

to significantly outperform conventional lipreading systems. The top 1 scores in Chung and Zisserman (2016) are less than Wand et al. (2016) and to date the best end-to-end system is that of Stafylakis and Tzimiropoulos who achieved an error rate of 11.29% on a 500 word vocabulary (Stafylakis and Tzimiropoulos, 2017b).

In developing lipreading systems we know that speech is a bimodal signal, and we use the visual channel of information for recognition of visual cues or gestures (Petridis et al., 2013). The units within this information channel, in sequence form a signal of its own, but it has no formal definition despite a variety of options presented previously (Cappelletta and Harte, 2011; Hilder et al., 2009; Chen and Rao, 1998; Fisher, 1968; Hazen et al., 2004). Irrespective of the definition in each paper, these units are commonly referred to as 'visemes' and in this paper, we define a viseme as a visual cue (sometimes also referred to as a gesture) that represents a subset of identical phonemes on the lips (Bear et al., 2015a; Nitchie, 1912; Bear, 2017). This means a set of visemes is always smaller than the set of phonemes (Cappelletta and Harte, 2012). These visemes are interesting because they help researchers to answer questions about how best to decipher lip motions when affected by issues such as human lipreading (Jeffers and Barley, 1971), language (Newman and Cox, 2012), expression (Metallinou et al., 2010), and camera parameters like resolution (Bear et al., 2014a).

Previous work has shown the benefits of deriving speaker-dependent visemes (Kricos and Lesner, 1982; Bear and Harvey, 2017) but the cost associated with generating these is significant. Indeed the work by Kricos and Lesner (1982) was limited due to the human subjects required, whereas the data-driven method of Bear and Harvey (2017) could scale if visual speech ground truths for the test speakers were available in advance. The concept of a unique Phoneme-to-Viseme (P2V) mapping for every speaker is daunting, so here we test the versatility and robustness of speaker-dependent visemes by using the algorithm in Bear et al. (2014b) to derive single-speaker, multi-speaker, and multi-speaker-independent visemes and use these in a controlled experiment to answer the following questions; To what extent are such visemes speaker-independent? What is the similarity between these sets of visemes?

This work is motivated by the many future applications of viseme knowledge. From improving both lipreading and audio-visual speech recognition systems for security and safety, to refereeing sports events and understanding silent films, understanding visual speech gestures has significant future impact on many areas of society.

In our previous work we investigated isolated word recognition from speaker-dependent visemes (Bear et al., 2015a). Here, we extend this to continuous speech. Benchmarked against speaker-dependent results, we experiment with speakers from both the AVLetters2 (AVL2) and Resource Management Audio-Visual (RMAV) datasets. The AVL2 dataset is a dataset of seven utterances per speaker reciting the alphabet. In RMAV the speakers utter continuous speech, sentences from three to six words long for up to 200 sentences each. Our hypothesis is that, with good speaker-specific visemes, we can negate the previous poor performance of speaker independent lipreading. This is because, particularly with continuous speech, information from language and grammar create longer sequences upon which classifiers can discriminate.

The rest of this paper is structured as follows: we discuss the issue of speaker identity in computer lipreading, how this can be a part of the feature extraction method to improve accuracy and how visemes can be generated. We then discuss speaker-independent systems before we introduce the experimental data and methods. We present results on isolated words and continuous speech data. We use the Wilcoxon signed rank (Wilcoxon, 1945) to measure the distances between the speaker-dependent P2V maps before drawing conclusions on the observations.

Please cite this article as: H. Bear, R. Harvey, Comparing heterogeneous visual gestures for measuring the diversity of visual speech signals, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.05.001