Image and Vision Computing 51 (2016) 1–12

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Visual units and confusion modelling for automatic lip-reading☆

Dominic Howell, Stephen Cox*, Barry Theobald

School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK

ARTICLE INFO

Article history:
Received 26 June 2015
Received in revised form 20 January 2016
Accepted 3 March 2016
Available online 1 April 2016

Keywords:
Lip-reading
Speech recognition
Visemes
Weighted finite state transducers
Confusion matrices
Confusion modelling

ABSTRACT

Automatic lip-reading (ALR) is a challenging task because the visual speech signal is known to be missing some important information, such as voicing. We propose an approach to ALR that acknowledges that this information is missing but assumes that it is substituted or deleted in a systematic way that can be modelled. We describe a system that learns such a model and then incorporates it into decoding, which is realised as a cascade of weighted finite-state transducers. Our results show a small but statistically significant improvement in recognition accuracy. We also investigate the issue of suitable visual units for ALR, and show that visemes are sub-optimal, not but because they introduce lexical ambiguity, but because the reduction in modelling units entailed by their use reduces accuracy.

1. Introduction

In the past thirty years, the development of automatic speech recognition (ASR) has received enormous attention to the point where ASR is now a useful and reliable technology. By contrast, automatic lip-reading (ALR) has received very little attention. This is not surprising, since lip-reading is used by only a very small proportion of the population who have hearing difficulties, and although some of these users can apparently lip-read with high accuracy, it is an imperfect form of communication. Audiovisual speech recognition (AVSR) is now gaining in importance as attention turns towards making ASR more robust to interfering noise. A number of different techniques have been proposed for AVSR, but all of them would benefit from higher accuracy when decoding speech purely from a visual signal. Although this is the most significant motivation for researching ALR, it also has a number of possible applications in its own right in areas such as provision of automatic training systems for teaching lip-reading, as an aid for people who are able to make speech gestures but whose voice function has been removed, and in fighting crime, as well as being an interesting topic in speech communication.

Speech is primarily an audio form of communication, and a considerable amount of information about speech sounds is missing from the visual speech signal [1]. The approach taken in this paper is to acknowledge that errors will occur in ALR because of this missing information, and to model and compensate for them, an approach which was inspired by previous work on dysarthric speech [2]. Dysarthric speakers have poor control over their articulators because of medical conditions (such as cerebral palsy, stroke, brain tumour etc.) that affect their motor functions. This leads to a reduced phonemic repertoire and poor quality articulation, and hence to speech that has low intelligibility and is difficult for ASR systems to recognise. Similarly, in visual speech, certain speech sounds cannot be distinguished because they differ in a feature that is not present in the visual signal (e.g. voicing, place of articulation when it is in the rear of the vocal tract). In previous work on dysarthric speech recognition, patterns of phonemic confusions made by a talker were learnt by the system, and when these confusions were compensated at recognition time, recognition accuracy increased [2]. In this work, we take a similar approach to lip-reading: we model visual speech as if it were a speech signal produced by a speaker who has a limited phonemic repertoire, and learn the resulting patterns of phoneme confusion by comparing the ground-truth phoneme sequences with the recognised sequences. At recognition time, we find the most likely interpretation (word-sequence) of the distorted phoneme output sequence in the light of these patterns. The approach is conveniently realised as a cascade of weighted finite-state transducers (WFSTs), one of which implements the confusion modelling, whilst the others implement familiar speech recognition tasks such as a pronunciation dictionary and language modelling. We compare this approach with

☆ This paper has been recommended for acceptance by Richard Bowden, PhD.

* Corresponding author.
E-mail addresses: Dominic.Howell@uea.ac.uk (D. Howell), s.j.cox@uea.ac.uk (S. Cox), B.Theobald@uea.ac.uk (B. Theobald).

http://dx.doi.org/10.1016/j.imavis.2016.03.003
0262-8856/

the standard speech recognition approach in which no knowledge of confusions is used.

Until recently, the ALR community has concentrated (with a few exceptions) on small and restricted lip-reading tasks, usually isolated letters and/or digits, as this kind of task is appropriate in the initial stages of developing a technology. Here, we report ALR results on continuous speech utterances that have a medium-size (~1000 words) vocabulary. We use a specially-recorded dataset consisting of videos of 3000 sentences spoken by a single speaker.

This unusually large corpus enables us to investigate a fundamental question in ALR, which is whether the use of phoneme-to-viseme mappings is effective. Visemes (discussed more thoroughly in Section 4) are claimed to be the visual equivalent of phonemes i.e. they are units of visual speech. It is common practice to employ a phoneme-to-viseme mapping (several are available) in ALR on the grounds that there are many phonemes that cannot be distinguished visually, and indistinguishable phonemes should logically be grouped together as a single unit for purposes of recognition. Although there has been some work on testing these mappings [3,4], it is not conclusive, and we investigate this in the first part of this paper.

The paper is organised as follows: in Section 2, we set the scene for our work by reviewing the state-of-the-art in ALR. Section 3 describes the two databases that we recorded for these experiments, and Section 4 describes our work in exploring the mapping between phonemes to visemes. Section 5 gives a brief background to WFSTs and describes our new approach in detail. Results based on the two databases used are described in Sections 6 and 7 respectively. We conclude with a discussion in Section 8.

2. Previous work

The first attempts to automatically recognise speech from a visual signal date back to the 1980s and the work of Petajan [5,6]. Even from that date, the focus was on using the visual signal to enhance audio ASR, and most work since then has concentrated on such integration rather than lip-reading per se. However, this work was important in laying the foundations for techniques of deriving features suitable for speech recognition from visual images. These early systems tended to use very small vocabularies, such as a subset of the alphabet or the ten digits, uttered by a single speaker [7,8], and used classification techniques such as hidden Markov models [9], neural networks [10] or hybrid models [11,12]. Work on continuous speech began about 2000 with continuously spoken digits [13]. A summer workshop at Johns Hopkins in 2000 [14] enabled major advances in AVSR by recording a very large database of 290 speakers speaking material with a vocabulary of 10,500 words (unfortunately it is unavailable). It pioneered the use of active appearance models (AAMs, [15]) as visual features and produced some of the first sets of speaker-independent ALR and AVSR results. Since then, there have been many different approaches to AVSR [16] including coupled HMMs [17], dynamic Bayesian networks [18], use of articulatory-based features [19], segment-based approaches [20,21] and more recently, deep neural networks [22,23]. A recent review of AVSR research that considers especially the selection of visual features for visual speech is [24].

Work in ALR itself has grown significantly in the last ten years, although many authors use the term "lip reading" to describe work in AVSR rather than ALR. The work has covered essentially three areas: development of new visual features [25-28], research into suitable units for lip reading [29-31] and exploration of new classification techniques [26,32,33]. Much of this work still uses small datasets of isolated words from a single speaker but a recent paper [34] presents speaker-independent results on a 1000 word connected speech task.

Table 1
Statistics of the ISO-211 and RM-3000 corpora.

	ISO-211	RM-3000
Total number of sentences	–	3000
Total number of unique words	211	979
Total number of unique phonemes	45	45
Total number of word tokens	1255	26,114
Total number of phoneme tokens	7040	105,561
Average number of words per sentence	–	8.70
Average number of phonemes per sentence	–	35.19
Average number of phonemes per word	5.61	4.04

3. Data and visual features

We recorded two datasets for the experiments in this work. A single speaker was recorded in each to eliminate the variation in visual features between speakers. We consider that this is a good strategy when exploring an innovative technique such as the one proposed here. In other recent work using multiple speakers from the large LiLiR dataset [35], we have shown how to compensate (to some extent) for speaker variation by using techniques such as speaker adaptive training and deep neural networks, and these techniques can be added later to the work described here.

The first dataset, called ISO-211, was an audio–visual database of 211 isolated words. It was designed for rapid experimentation in developing WFSTs for lip-reading. ISO-211 has a vocabulary of 211 phonetically rich words which were chosen to give maximum bigram coverage. The data were captured in a specialised recording environment using a Sanyo Xacti camera in portrait orientation at 1080 × 1920 pixel resolution using progressive scan at a sampling frequency of 59.94 frames per second. Audio was captured using a clip microphone at a sampling frequency of 48 kHz. A single native English speaking female speaker spoke six repetitions of each word.

The second dataset, called RM-3000, consists of audio–visual recordings of 3000 sentences spoken by a single native English-speaking male speaker. The sentences were randomly selected from the 8000 sentences in the Resource Management (RM) Corpus [36]. The motivation for recording RM-3000 was to obtain a large database of continuous visual speech that had a medium size vocabulary and that was spoken by a single speaker. Sentences from the RM Corpus were chosen because its format (sentences of varying length whose grammar can be well-modelled with a language model) and its vocabulary size (1000 words) are ideal for research into lip-reading in its current state of development. The recording setup was the same as for the ISO-211 dataset.

Phoneme transcriptions of the sentences were derived from the BEEP Dictionary [37]. Some statistics about the two databases are shown in Table 1.

3.1. Features for lip-reading

In [38], three video resolutions (640 × 360, 1080 × 720 and 1920 × 1080) were compared in a visual-phone lip-reading recognition task, and it was found that there was no significant difference in the accuracy obtained. Therefore, to improve the efficiency of the feature extraction and modelling processes, all videos were down-sampled to a third of their original resolution to 360 × 640 pixels. Between 20 and 30 frames from each recording session were selected for hand-labelling: we labelled frames that described the extremities of mouth movements to capture as much variance of shape and appearance possibilities as possible. In each selected frame, 111 points were labelled over the whole face to ensure stability when