Survey on Automatic Lip-Reading in the Era of Deep Learning

Adriana Fernandez-Lopez, Federico Sukno

PII: S0262-8856(18)30127-6
DOI: doi:10.1016/j.imavis.2018.07.002
Reference: IMAVIS 3707

To appear in: *Image and Vision Computing*

Received date: 10 April 2018
Accepted date: 14 July 2018

Please cite this article as: Adriana Fernandez-Lopez, Federico Sukno , Survey on Automatic Lip-Reading in the Era of Deep Learning. Imavis (2018), doi:10.1016/j.imavis.2018.07.002

Survey on Automatic Lip-Reading in the Era of Deep Learning

Adriana Fernandez-Lopez and Federico Sukno

Department of Information and Communication Technologies,
University Pompeu Fabra, Barcelona, Spain.

Abstract

In the last few years, there has been an increasing interest in developing systems for Automatic Lip-Reading (ALR). Similarly to other computer vision applications, methods based on Deep Learning (DL) have become very popular and have permitted to substantially push forward the achievable performance. In this survey, we review ALR research during the last decade, highlighting the progression from approaches previous to DL (which we refer to as traditional) toward end-to-end DL architectures. We provide a comprehensive list of the audio-visual databases available for lip-reading, describing what tasks they can be used for, their popularity and their most important characteristics, such as the number of speakers, vocabulary size, recording settings and total duration. In correspondence with the shift toward DL, we show that there is a clear tendency toward large-scale datasets targeting realistic application settings and large numbers of samples per class. On the other hand, we summarize, discuss and compare the different ALR systems proposed in the last decade, separately considering traditional and DL approaches. We address a quantitative analysis of the different systems by organizing them in terms of the task that they target (e.g. recognition of letters or digits and words or sentences) and comparing their reported performance in the most commonly used datasets. As a result, we find that DL architectures perform similarly to traditional ones for simpler tasks but report significant improvements in more complex tasks, such as word or sentence recognition, with up to 40% improvement in word recognition rates. Hence, we provide a detailed description of the available ALR systems based on end-to-end DL architectures and identify a tendency to focus on the modeling of temporal context as the key to advance the field. Such modeling is dominated by recurrent neural networks due to their ability to retain context at multiple scales (e.g. short- and long-term information). In this sense, current efforts tend toward techniques that allow a more comprehensive modeling and interpretability of the retained context.

Keywords: Automatic lip-reading, audio-visual corpora, visual speech decoding, deep learning systems, multi-view lip-reading.

1. Introduction

Speech is the most used communication method between humans, and it is considered a multi-sensory process that involves perception of both acoustic and visual cues. McGurk and McDonald demonstrated the influence of vision in speech perception in [1], where it was experimentally shown that when observers were presented with mismatched auditory and visual cues, they perceived a different sound from those presented in the stimulus, i.e. the syllable /ba/ was spoken over the lip movements of /ga/, and the perception was the intermediate syllable /da/. Since then, many authors have demonstrated that the use of visual information in speech recognition improves robustness [2, 3].

Despite audio signals are in general much more informative than video signals, it has been demonstrated that most people use lip-reading cues to understand speech. However, these cues are often used unconsciously and to different degrees depending on aspects such as the hearing capability [4] or the acoustic conditions (e.g. the visual channel becomes more important in noisy environments) [5], [6], [7], [8]. Furthermore, the visual channel is the only source of information for people with hearing disabilities to understand the oral language [9], [2], [10].

In the literature, much of the research has focused on Automatic Speech Recognition (ASR) systems, given

Email address: adriana.fernandez@upf.edu, federico.sukno@upf.edu (Adriana Fernandez-Lopez and Federico Sukno)