Audio-Visual Speech Emotion Recognition

A

Oryina Kingsley Akputu
Sunway University, Malaysia

Kah Phooi Seng
Edith Cowan University, Australia

Yun Li Lee
Sunway University, Malaysia

INTRODUCTION

One widely accepted prediction is that computing will soon move to the backgrounds, waving into tools of everyday living, and projecting human users in the foregrounds. Henceforth the future affective computing environments will need to focus more on human-centred designs than computer-cantered ones. Current human computer interaction(HCI) designs involves interfacing devices such as mouse and keyboard, which transmit only explicit signals while ignoring implicit information like emotions from users. However, tailoring such designs to user's emotional expressions has potentials for enhancing mundane application in intelligent communications systems. One specific application is found in judicial courtrooms(Fersini, Messina, & Archetti, 2012), where salient portion in a multimedia clips of relevant debate sessions are analysed for reviewing noticeable affective state of a subject. Another applicable area for this type of technology is in affective context acquisition using audio-visual signals for Ambient Intelligence(Schmalenstroeer & Haeb-Umbach, 2010). The vision for this type of technology, describes systems that are, embedded in our surrounding, present when needed, think on their own and can make our lives better. However, achieving these as expected will require equipping machines with sufficient intelligence to recognize human speech with high accuracies.

Nevertheless attaining high level intelligence with substantial degree of accuracy under less constrained settings remains a challenge due to the following reason. The acoustic variability introduced by different sentences, speakers speech styles and rates introduces a major obstacle as it directly affect common extracted speech features including pitch and energy contours. Consequently tremendous research attempts are made for audio-visual speech emotion recognition tasks, which refers to the process of converting human speech into a sequence of emotion annotated words. This article presents a comprehensive overview of audio-visual speech emotion recognition systems including key techniques and findings from previous researches up to date. Key component tasks including feature extraction with classification frameworks used are extensively discussed. This then serves as a basis to unveiled various challenges for future research directions. The article targets infant pattern recognition researchers who may need a coincided tutorial for a basic background in emotional speech analysis. Moreover an advance researcher may use this article for a precised reference too.

BACKGROUND

Since the early 19th century, research works have been conducted in search for best models for analysing and interpreting subtlety complex human emotions based on the following common dimensions (Scherer, 2000): *Continuous abstract dimensions*, *appraisal dimensions* and *discrete categories*. In continuous dimension space, a multidimensional space is defined in which emotions categories are represented by underlying dimensional points such as; Valence(V), Arousal(A), and control (Dominance) (C).In appraisal dimensions, emotional process and their related events are described. The central idea is to specify a set of criterion whose speculation underlies emotional constituents of an appraisal

DOI: 10.4018/978-1-4666-5888-2.ch011

process. The discrete categories on the other hand entail selecting a set of word labels for representing emotions. In computer vision, and HCI the prototypical (archetypical) or discrete emotions includes, Anger, Disgust, Fear, Joy, Sadness and surprise(Cowie et al., 2001).

Nevertheless emotional expression drawn from any of the above dimensions, are known for their multimodal correlation and complexity. Traditionally, researchers have either employed, single modality or multimodal approach in the task of audio-visual emotion recognition. Information processed by a single sensor (modality) is limited to a single sensory cue. For instance, utilizing facial expression videos or audio-signal of an utterance separately for emotion recognition. Multimodal speech approaches however combine effective cues from audio and visual signals. Integrating audio and visual signals is introduced with the sole aim of harnessing individual advantages inherent in underlying modalities. A more basic audio-visual speech emotion recognition system is composed of four components: audio feature extraction, visual feature extraction, feature selection and classification. What may be considered the structure of a standard audio-visual emotion recognition system is illustrated in Figure 1. More crucial among components of this structure shall be discussed in the following sections of this article.

Figure 1. Standard audio-visual speech emotion recognition system