Robust speech processing using multi-sensor multi-source

information fusion--an overview of the state of the art

Parham Aarabi a,\*, Belur V. Dasarathy b a Electrical and Computer Engineering, University of Toronto, 10 Kings College Road, Toronto, ON, Canada M5S3G4 b Huntsville, AL, USA Received 3 December 2003; received in revised form 3 February 2004; accepted 3 February 2004

Abstract

This article offers an overview of the state of the art in robust speech processing and delineates the role of information fusion in furthering its objectives. In addition, it also serves the function of the traditional guest editorial of a special issue in terms of presenting a brief introduction to its contents. 2004 Elsevier B.V. All rights reserved.

1. Introduction

This article offers an overview of the state of the art in robust speech processing and delineates the role of information fusion in furthering its objectives. In addition, it also serves the function of the traditional guest editorial of a special issue in terms of presenting a brief introduction to its contents.

@2004 Elsevier B.V. All rights reserved.

Keywords: Robust speech processing; Multisensor information fusion; Speech recognition

For decades, extensive research has been conducted on enabling computers to process (i.e. localize, enhance, and/or recognize) speech [1–4,16–20,33]. During this time, numerous advancements in automatic speech recognition and processing have been made, resulting in various commercial speech recognition systems. These state-of-the-art speech recognition systems often have a high recognition accuracy rate obtained through extensive training and computations. The high accuracy rate, however, substantially decreases in the presence of noise, such as secondary conversations. As a consequence, robust speech recognition in noisy environments continues to be a popular research topic as a yet unresolved challenge [16,18,20,21]. Recent advancements in single-microphone speech recognition (including probabilistic speech enhancements and noise modeling techniques such as [16,21]) have gone a long way in addressing some of the robustness issues with current speech recognition systems. There is, however, a limit as to what can be achieved with a single microphone. Systems that employ multiple sensors (including cameras and microphones) have a potential for greater robustness and higher recognition accuracy rates. The fusion of multiple sensors or multiple information sources for robust speech processing is the focus of this special issue.

2. Literature review

The application of information fusion methodologies for speech processing applications has been explored in the past [4,6,8]. For example, in [4], an audio-visual vowel recognition system was explored and analyzed using an elucidative fusion paradigm. This work not only illustrated the benefits of the fused multi-modal system, but also illustrated the relative influence of the individual modalities on the fused result. For example, at a signal-to-noise ratio of )24 dB, the visual modality had a much larger relative influence over the acoustic modality for the correct decisions than it did when there was no noise present. This clearly illustrates the benefits of information fusion: in different situations, some information sources may be unreliable, and as long as a proper method for the fusion of these sources exists, the relative influence of the unreliable sources will be smaller than those of the more reliable sources. Similar work in audio-visual supports this claim [4–8,23,34–39]. Audio-visual information fusion has for many years been an area of extensive research. In [25], a hybrid audio-visual feature fusion and audio-visual decision fusion methodology is employed for robust speech recognition. In [26], a coupled hidden Markov model (which is a special case of Dynamic Belief Networks) is employed for audio-visual information fusion for speech recognition. In [27], a weighted decision fusion approach is employed for audio-visual speech recognition. A variety of other audio-visual speech recognition systems, which fuse the signal of a microphone with the images recorded by a camera, have also been proposed [28–32]. In all of these systems, the fused audio-visual system results in a substantial reduction in the word-error-rate as compared to either the visual or the acoustic modality by itself.

In spite of all of this research, fusing the information obtained by multiple sensors, some of which could be of different modalities, is still an open-ended and by its very nature a vaguely defined problem [2,8,22]. First, there is \the question as to exactly what type of sensors should be used. Combining cameras and microphones, for example, can result in greater robustness since cameras are unaffected by acoustic noise and microphones are unaffected by events (such as objects, obstacles, lighting, etc.) that affect cameras. However, fusing different modalities is often more difficult and complex, which can lead to greater computational requirements. Nevertheless, even with the existence of this complexity-effectiveness trade-off, numerous systems have attempted to fuse multi-modal information for a variety of applications. Examples include audio-visual speech recognition systems employing a single camera and a microphone, resulting in a higher speech recognition accuracy rate, and a greater robustness to noise [5–15]. Other applications include audio-visual sound localization [2,3], where a speaker is localized visually using multiple cameras and acoustically using multiple microphones. As expected, these systems illustrated improved performance over the audio-only or the video-only localization system. The next issue that arises is the point at which the information is fused. For example, the signal values that are obtained by a set of microphones could be integrated at the signal level (using techniques such as beamforming), or, they could be integrated after speech recognition is performed on each channel. In fact, this integration or fusion can occur anywhere in the speech recognition or localization process. Most of the prior work in this area has focused on data fusion prior to the recognition process (i.e. the signals are integrated to result in a cleaner signal which is then used for recognition). Such a pre-recognition fusion has the advantage of computational simplicity, although its optimality is not very clear. Pre-recognition data fusion can be effective when an array of microphones is used. When cameras and microphones are used jointly, it is far more difficult to combine their data at the signal level. Any such fusion would have to involve a projection of the visual results into the space of acoustic results, or vice versa. Clearly, this projection can be a very complex and computationally demanding task. Nevertheless, there have been attempts at early-stage audio-visual information fusion [1,11,12] as well as late-stage audio-visual information fusion [14,15]. These examples explore the relative tradeoff between increased complexity and increased speech recognition accuracy.

So, the questions at hand are (1) what do we fuse, (2) when (i.e. at what stage of the recognition process) do we perform this fusion, and (3) how do we actually perform the fusion. These are three inter-related questions whose final answers may be years away. Nevertheless, the six papers in this special issue provide some unique and innovative solutions to these questions.

3.Special issue overview

The first paper, which introduces asynchronous hidden Markov models (HMMs) for audio-visual speech recognition, proposes an asynchronous HMM structure, which is jointly applied to acoustic and visual data obtained from a microphone and camera, respectively. It is shown that the proposed fusion model outperforms the standard HMM architecture for either audio-visual speech recognition or for acoustic-only recognition, especially at lower signal-to-noise ratios (SNRs).

The second paper also considers the audio-visual information fusion problem by performing decision fusion for continuous audio-visual digit recognition. Their approach involves a post-recognition fusion (decision fusion) of the recognition results obtained from the visual data and the results obtained from the acoustic data. It is shown that at very low SNRs (15 dB), their system performs similar to the audio-only system and far better than the visual-only system. In the mid-SNR range (0–15 dB), their system outperforms both the audio-only and the video-only systems.

The third paper in this special issue attempts to fuse two microphones and two cameras at the signal level for speech enhancement. Here, the visual data that are related to speech (i.e. the lip movements) are initially translated into two corresponding acoustic signals. These visually derived or virtual acoustic signals are then used as a basis for speech separation using the signals obtained from the two microphones

The fourth paper deals with the very interesting topic of quantizing the worth of a sensor in a multi-sensor setting. In a setting were multiple sensors such as microphones are present, the utilization of all sensors for speech processing applications may not be needed. However, in order to determine what sensors should be utilized, the amount of unique information provided by that sensor must be measured. This study proposes an information theoretic measure for exactly this purpose.

The fifth paper in this special issue employs an array of 24 microphones for the control and navigation of a robotic tour-guide. Here, the data collected from the 24 microphones are processed in order to estimate the location of the speech source, which is assumed to be the robot. This multi-sensor acoustic localization is used to correct the position of the robot and to steer it towards its next destination. It is interesting to compare this fifth paper with the system proposed by Enzo et al. [24]. Unlike the technique of [24], where a microphone array onboard the robot was used for human operator localization and for improved human–robot interactions, this study utilizes a fixed array placed in the environment. It is interesting to note that the proposed acoustic robot localization systems are in fact complementary, making the fusion of the two an interesting and possibly fruitful direction of future work.

The sixth paper, which focuses on feature fusion for robust speech recognition, proposes a novel method of improving speech recognition accuracy rates. Instead of relying on multiple microphones or audio-visual information fusion, they rely upon two information sources consisting of the acoustic features (the Mel-frequency cepstral coefficients) and the articulatory features.

4. Concluding comments

The ultimate goal of realizing a robust and accurate speech recognition system requires extensive research into computationally efficient methods to localize, enhance, and recognize speech. Just as we humans use our two ears and two eyes for speech localization and recognition, artificial speech recognition systems will also require multi-modal and multi-sensor information fusion in order to achieve accurate results in practical situations. As such, the realization of such systems requires extensive research on fusing the information obtained from multiple microphones and cameras, as well as other sensors/sources. The six papers in this special issue provide some rather unique and positive steps towards this ultimate goal; an ultimate goal which may one day enable robust speech recognition to be used in cars, appliances, tablet computers, portable computers, personal digital assistants, and for the disabled.