A Multi-modal Perceptual System for Social Robots: A Brain-inspired Architecture

by

Mohammad Al-Qaderi

M.Sc., Jordan University of Science and Technology, 2007

B.Sc., Jordan University of Science and Technology, 2002

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
School of Mechatronic Systems Engineering
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Fall 2018

Approval

Name:	**Mohammad Al-Qaderi**
Degree:	**Doctor of Philosophy**
Title:	**A Multi-modal Perceptual System for Social Robots: A Brain-inspired Architecture**
Examining Committee:	**Chair:** Shahram Payandeh Professor

Ahmad Rad

Senior Supervisor
Professor

Mehrad Moallem

Supervisor
Professor

Ed Park

Supervisor
Professor

Carlo Menon

Internal Examiner
Professor

Goldie Najat

External Examiner
Associate Professor
Department of Mechanical & Industrial Engineering
University of Toronto

Date Defended/Approved: November 2, 2018

Abstract

The main theme of this research study is machine perception. We are particularly interested in developing a perceptual system for social robots. These robots are designed to communicate with humans the same way they interact with each other. We argue that in order to meet this stern criterion, it is sensible that such robots are capable of perceiving their environment in a similar fashion to humans. The thesis focusses on developing a framework for designing a human-oriented perceptual system for social robots. The research is intrinsically interdisciplinary and requires integration of ideas from psychology, psychophysics, and neuroscience about human perception with robotics engineering.

First, the skeleton of the architecture is developed motivated by the understanding of the hierarchical structure of primate sensory cortex. The key sub-systems of the architecture and interrelationship among them are shaped by insights from biological, computational, and psychological understanding of human perception. In particular, the multi-modal sensory information processing in the sensory cortex, the spatial-temporal binding criteria, and limited human's channel capacity of processing information. The system encapsulates the parallel distributed processing of real-world stimuli through several sensor modalities and encoding them into features vectors which in turn are processed via a number of dedicated processing units through hierarchical paths. The proposed perceptual system is context independent and can be applied to many on-going problems in social robotics.

Next, a customized version of the system is developed to address the problem of person recognition in social settings. The system utilizes the information from visual and auditory modalities via a non-invasive methodology as opposed to reported person recognition systems that generally invasive. We adopt spiking neural network to integrate information form the available sensor modalities to provide a plausible and realistic computational model that facilitate a real-time response in various challenging scenarios of person recognition in social settings.

In the last stage, a robust speaker recognition system has been designed in the light of the above framework. The system exploits prosodic feature to reduce the population size as well as integrates the advantages of multi-feature system and multi-classifiers system to overcome the challenges of speaker recognition in noisy environments and availability of only short utterances.

Keywords: human-oriented perceptual system; brain-inspired approach, spiking neural networks, multi-modal person recognition, speaker recognition.

To my parents, my wife, and my lovely twin daughters

Acknowledgements

In the name of Allah, the Merciful, the Compassionate

First and foremost, my deep gratitude to Almighty Allah, all the praise and glory are to Him for giving me the patience, strength, health, time, resources and opportunity to complete my Ph.D. thesis.

I would like to express my sincere and deep gratitude to my senior supervisor, Dr. Ahmad Rad, for his continuous support and invaluable guidance throughout my Ph.D. studies. I would like to thank him for providing me with the independence to freely explorer this research project while providing thoughtful feedback and constructive criticism to stay on track. I am indebted to him to introduce me to this interdisciplinary approach of machine perception.

I would also like to express my thanks to the members of my supervisory committee, Dr. Mehrdad Moallem and Dr. Edward Park for their time reading this thesis and for providing insightful comments. Also, I am grateful to Dr. Shahram Payandeh for being the defense committee chairman. Thanks also to my internal examiner Dr. Carlo Menon and my external examiner Dr. Goldie Nejat for their participation in my dissertation committee and their valuable comments and advice.

I feel privileged to have had the opportunity to work with so many inspiring colleagues in AISL Lab, whose support, friendship, and encouragement making the whole journey worthwhile.

Most importantly, I am eternally grateful to my parents for their never-ending support, love, and patience, without which this would not have been possible. Heartfelt thanks to my beloved wife, Noor, this thesis could not have been completed without her love, support, and unwavering belief in me. I can not find words to express the love and gratitude I have for her.

Table of Contents

List of Figures

List of Acronyms

AI	Artificial Intelligence
LIF	Leaky Integrate-and-Fire
SLAM	Simultaneous Localization and Mapping
HRI	Human-Robot Interaction
SEAI	Social Emotional Artificial Intelligence
SNS	Synthetic Nervous System
DAC	Distributed Adaptive Control
FACE	Facial Automaton for Conveying Emotions
DPU	Dedicated Processing Unit
AIT	Anterior Inferotemporal Cortex
FIT	Feature Integration Theory
CNN	Convolutional Neural Networks
DCNN	Deep Convolutional Neural Networks
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
RNN	Recurrent Neural Network
MFCC	Mel-Frequency Cepstral Coefficients
PLP	Perceptual Linear Prediction
GFCC	Gammatone Frequency Cepstral Coefficients

Chapter 1.

Introduction

1.1. Background and motivation

On the amazing flexibility of the human perceptual system, Anne Treisman (1986) wrote "*Just as reading is "externally guided thinking" (Neisser, 1967, p. 136), so perception may be a form of controlled hallucination"* [1]. Thirty-two years on since that lucid encapsulation of the essence of the human perception process and at the time when we know a lot more about the mind and the intricate instrument of brain; our understanding of the human perceptual system is still opaque. Nonetheless, researchers in neuroscience, psychophysics, and psychology have suggested several plausible and some experimentally verifiable models of the human perception. The so-called "sense-data" model of the perception [2], also referred to as "naïve" perceptual system, is widely known. Antonio Damasio, the acclaimed neurobiologist, suggested a multi-modal theory of perception centered on the notion of "cell assemblies" and "convergence zone" [3]. Whereas each sensory modality (sight, sound, smell, touch) has its own dedicated unimodal pathway (unimodal cortex), an excitation triggers simultaneously (or nearly simultaneously) a response in more than one sensory modality. For example, when we hear a voice, the superior colliculus area of the brain is responsible to process both auditory and visual stimuli (we turn around towards the direction that the sound comes from).

Inspired by the extraordinary architecture of the human multimodal perceptual system and the curiosity to search for an answer to the challenging research question of "*can a similar, although vastly simplified, architecture be designed for social robots?*" were the motivation behind this research project. The rationale was if social robots were going to be employed in human social settings (homes, offices, hospitals, etc.), should they be equipped with a perceptual system that has a one-to-one correspondence (analogy) to our own perceptual system? Should they function within the human reaction time? Is the presence of all modalities required for the system to work (efficiently)? Is it possible to devise a context independent perceptual system that can be applied to different applications (person recognition, emotion recognition, etc.)? How the architecture of such a system would look like? The attempt to answer the above research questions required delving into a complex interdisciplinary project – not a trivial task for an engineer with limited training outside the discipline!

It is conceivable that towards the midst of this century and beyond, intelligent life-like androids and cyborg-type systems will share the environment with humans. These machines, hereafter referred to as social robots, will be complex and best fit within the systems of systems architecture. Social robots will likely be patterned after humans and are expected to function predominantly in an environment designed for humans. These robots are primitive versions of the future "*humanized robots*" i.e., the character Ava in Ex Machina. Not getting too much ahead of us, the basic versions are already available (Nao, Pepper, etc.) and are envisaged to being employed in diverse applications including but not limited to elderly and disabled caregiver and companion, health care, education, film and entertainment, and museum tour-guide to name a few [4]–[13]. Social robots may have the same core and native abilities such as navigation (SLAM, path planning, etc.), learning (AI) capabilities, and advanced sensors with their counterparts (service / field robots, driverless cars); however, they have additional attributes including recognition of human emotions as well as human-friendly and natural communication skills. Essentially, these robots belong to special class of intelligent and autonomous robots that are specifically designed to work with humans and in human social settings. Social robots are normally deployed in environments designed for humans and not specifically structured for robots. They are also expected to interact with humans in diverse social settings in a natural manner (the way humans interact among themselves), and are designed to be socially acceptable to humans in terms of their appearance and the way they communicate (with humans).

The central motivation of this research project is to address the less studied problem of robotic perception as opposed to well-studied areas of robotics sensors, navigation, and affective computing. The abundance and variety of sensors and relevant technologies have inadvertently masked out the importance of the underlying perception system. Here we argue that sensors, important as they are, produce only data and not information. It is plausible then to suggest that research into developing perceptual systems should be given at least the same priority as focusing on data from sensor networks and subsequently attempting to disentangle its by-product as a new problem of "big data" analysis. Within this context, the seamless integration of social robots into human social settings and their ultimate acceptance by humans will largely depend on how natural (in human sense) such robots behave. Research by Niculescu et al. in human-robot interaction suggested that people considered the level of closeness to human-like response from a social robot a highly favorable attribute [14]. The authors argued that the real-time response at the level of human rate was more important than the accuracy and the correctness of the response itself. In related studies [15]–[17], authors point to the importance of human-oriented perception, reading and expressing social cues, and real-time response at human interaction rates as important parameters that must be addressed by social robotic systems. These machines must be capable of sensing, perceiving, and interpreting the real-world environment similar or as close as possible to that of humans. Therefore, mimicking the way humans process stimuli, and synthesizing similar interpretations of those stimuli as of humans should be among the salient features of social robotic systems.

In this thesis, we report an end-to-end perceptual system specifically designed for social robots as we argue that these robots require an underlying perceptual system rather than an ad-hoc interface of different sensors and computational methods to solve a specific problem like face recognition. The proposed architecture is a potential solution to the following questions: what are the most distinctive features of real-world environments that should be considered in the context of social robot's objectives? what aspect of measures should be employed to address these features? and how does the perceptual system processes and interprets the data provided by the sensory system? The designed framework provides feasible answers to such questions. We present a modular parallel-distributed processing system that is inspired from the way that human's brain processes and perceives real-world environments.

1.2. Research outline

The research project was completed in four interrelated phases as shown Figure 1.1. In the first phase, motivated by findings in neuroscience, psychology, and psychophysics, we propose an architecture for a context independent multi-modal perceptual system specifically designed for social robots. We also include some design guidelines to customize the system for specific applications. Moreover, the top-down influences and temporal binding with fading memory were incorporated in the design of the perceptual architecture as a means of reducing the search scope and integrating the information from multisensory processing paths, respectively. In the second phase, a customized version of the perceptual system was developed and tested for the problem of person recognition in social settings. In this system, we considered diverse real-world scenarios of social human-robot interactions. We adopted the leaky integrate-and-fire neuron (LIF) model, which has an adjustable threshold value and can be modified to compromise between the reliability of the perception outcome and the time required to finalize the perception process, to integrate information from several sensor modalities. Extensive experimental studies were performed via applying real-world databases to evaluate the performance of the perceptual system in various scenarios of social human-robot interactions. In the third phase, the design guidelines have been utilized to develop a sophisticated speaker recognition system by incorporating the concept of top-down influences and the integration of multi-feature and multi-classifier in an elegant architecture to address challenging scenarios such as speaker recognition in noisy environments and with only access to short utterances. In the last phase, the design of an in-house multi-modal social robotics platform to test and demonstrate the properties of the perceptual system was realized. Real-time implementation of the top-down influences and temporal-binding with fading memory were evaluated in an in-house designed multi-modal social robot platform.

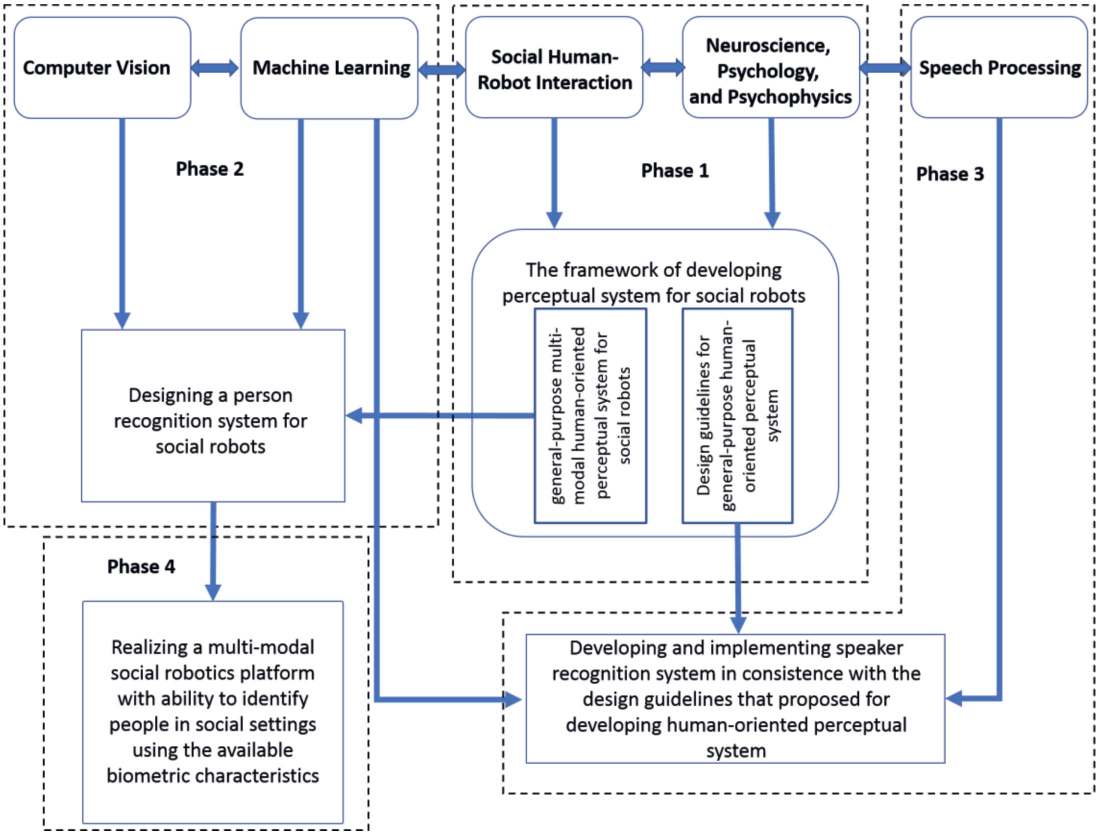

Computer Vision
Machine Learning
Social Human-Robot Interaction
Neuroscience, Psychology, and Psychophysics
Speech Processing
Phase 2
Phase 1
Phase 3
The framework of developing perceptual system for social robots
general-purpose multi-modal human-oriented perceptual system for social robots
Design guidelines for general-purpose human-oriented perceptual system
Designing a person recognition system for social robots
Phase 4
Realizing a multi-modal social robotics platform with ability to identify people in social settings using the available biometric characteristics
Developing and implementing speaker recognition system in consistence with the design guidelines that proposed for developing human-oriented perceptual system

Figure 1.1. The design flow of this research project

1.3. Organization of the thesis

This thesis is comprised of four chapters organized as follows:

Chapter 1 aims to set the scene along with background and motivation as well as providing a research road map for this research project.

In chapter 2, a selective literature survey on pertinent theories and principles will be presented. Moreover, relevant studies on social human-robot interaction and school of thoughts on designing autonomous robot are included. The chapter establishes the link between the aforementioned research areas in order to highlight the characteristics of an efficient perceptual system designed for social robots.

Chapter 3 provides a summary of the main contributions of the studies undertaken in this research project. A general-context multi-modal perceptual system for social robots a long with its key sub-systems and the interrelationship among them are discussed in

Sec.3.1. In Sec. 3.2, a solution to the problem of person recognition in social settings via a customized version of the proposed multi-modal perceptual system is also presented. In Sec. 3.3, the effect of incorporating top-down influences on the performance of the proposed multi-modal perceptual system is demonstrated in real-time implementation using an in-house designed social robotics platform. Finally, a speaker recognition system that developed in the light of the design guidelines for human-oriented perceptual system is introduced in Sec. 3.4.

Chapter 4 concludes with the summery of achievement of this thesis along with suggestions for future work.

1.4. Statement of originality

The main contributions and developments made by the author of this PhD thesis are summarized in the following statements:

- Designing a large-scale end-to-end social robotic perceptual architecture as a general-purpose multi-modal perceptual system for social robots.
- Suggesting two approaches for integrating the outputs of multisensory processing paths via temporal binding with fading memory and on-fly fuzzy inference system.
- Introducing top-down influences on multisensory information processing as a key to reduce the computational cost of the proposed architecture and to limit the search scope for stimulus candidate and hence facilitating social human-robot interaction.
- Synthesizing the key findings in neuroscience, psychology, and psychophysics about human perceptual process as design guidelines for a general-purpose human-oriented perceptual system.
- Realizing a social robotics platform that can be used as in various social human-robot interaction scenarios.
- Developing and implementing person recognition system for social robots inspired from the way that human identifies individuals during everyday life activities. This system uses the available biometric characteristics to identify an individual at the human interaction performance’s level and rate.
- Developing and implementing speaker recognition system in consistence with the design guidelines that proposed for developing human-oriented perceptual system.

1.5. Publications

At the time of writing this thesis, two journal papers have been published. Also, there is one paper that has been submitted to an international journal. The list of publications is as follows:

1. M. K. Al-Qaderi and A. B. Rad, "A Brain-inspired Multi-modal Perceptual System for Social Robots: An Experimental Realization," IEEE Access, vol. 6, pp. 35402–35424, 2018.

2. M. Al-Qaderi and A. Rad, "A Multi-Modal Person Recognition System for Social Robots," Appl. Sci., vol. 8, no. 3, p. 387, 2018.

3. M. Al-Qaderi, E. S. Lahamer, and A. Rad, "A Two-stage Classifier Speaker Recognition System Based on Prosodic and Spectral Features via Fuzzy Inference Fusion," Eurasip J. Audio, Speech, Music Process., (under review).

Chapter 2.
Literature Review

2.1. Introduction

Humans socialize and communicate with each other via language and other senses. The idea of a similar form of human-robot interaction has been suggested in numerous science fictions; however, with the rapid advancement in AI and affective computing in the last two decades, the realization of “natural” human-robot interaction is feasible. Social robotics, a subclass of autonomous robotics, is an interdisciplinary research area focusing on developing robots that could interact with humans within their acceptable social, emotional, and cultural norms [18], [19]. The central characteristic of this area of robotics is the closeness of the robot’s *artificial behaviour* to the human’s interactions among themselves. A social robot could share many attributes with an autonomous robot including autonomous navigation, path planning, learning capabilities, computational power, and on-board sophisticated communication system [19], [20]. Its perceptual and affective systems along with a human-like silhouette are, however, among its differences with an autonomous robot [21], [22]. We acknowledge, though, that some researchers do not think that a human-like form is a required feature of a social robot and put more emphasis on the interactive and affective competence. Much of the recent literature on social robotics is focused on the so-called affective dimension – recognising human’s emotional states and reacting appropriately. In this thesis, we suggest that whereas such features are regarded as the precursor for social robots; a sophisticated perceptual architecture, upon which the robot functions and interacts with humans and the world, is a gap in the state-of-art and is worth of investigation. Among numerous challenges, however, are how to develop and design perceptual, cognitive, and emotional systems within a unified infrastructure? Researchers adopt an interdisciplinary approach that bridges robotics engineering with neuroscience, psychology, psychophysics, and social sciences to facilitate designing these intelligent systems. Figure 2.1 depicts the main active research themes that represent the core of social robotics.

In this chapter, we attempt a succinct yet selective review of the relevant literature to this research project. Developing an efficient perceptual system for social robots is an extremely challenging undertaking due to its inherent complexities. It is an interdisciplinary endeavour molding diverse engineering research fields (computer vision, speech processing, AI, and machine learning, etc.) with the amalgamation of latest know-how in neuroscience, psychology, and social sciences. As such, the main purpose of this chapter is to set the scene and present the scope of this research and to establish its relation to the state-of-art. Henceforth, the emphasis in this chapter is directed towards presenting a broad yet focused review of the key relevant studies, as specific literature will be cited in the respective chapters of the thesis.

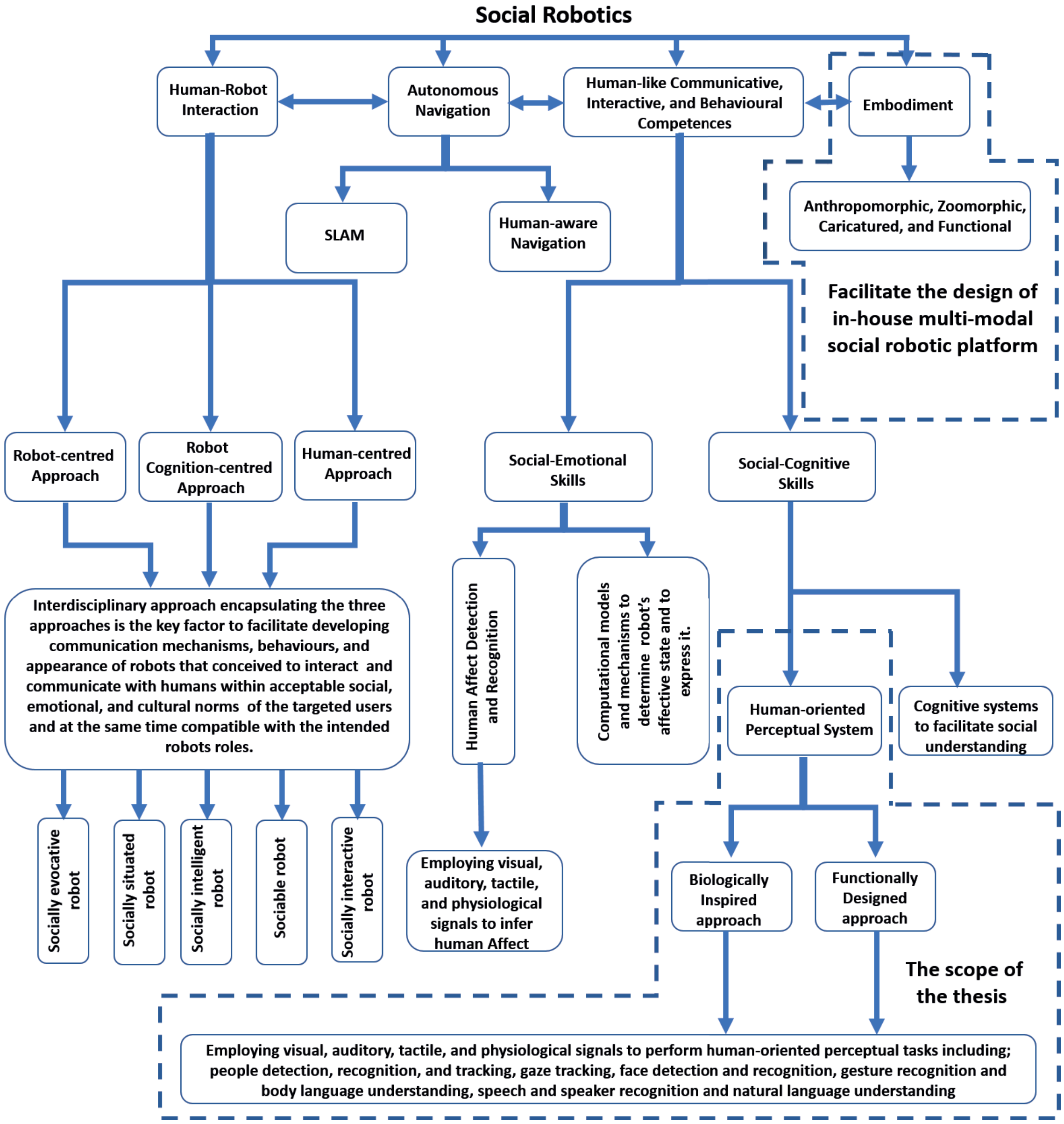

Social Robotics
Human-Robot Interaction
Autonomous Navigation
Human-like Communicative, Interactive, and Behavioural Competences
Embodiment
Anthropomorphic, Zoomorphic, Caricatured, and Functional
SLAM
Human-aware Navigation
Facilitate the design of in-house multi-modal social robotic platform
Robot-centred Approach
Robot Cognition-centred Approach
Human-centred Approach
Social-Emotional Skills
Social-Cognitive Skills
Interdisciplinary approach encapsulating the three approaches is the key factor to facilitate developing communication mechanisms, behaviours, and appearance of robots that conceived to interact and communicate with humans within acceptable social, emotional, and cultural norms of the targeted users and at the same time compatible with the intended robots roles.
Human Affect Detection and Recognition
Computational models and mechanisms to determine robot's affective state and to express it.
Human-oriented Perceptual System
Cognitive systems to facilitate social understanding
Socially evocative robot
Socially situated robot
Socially intelligent robot
Sociable robot
Socially interactive robot
Employing visual, auditory, tactile, and physiological signals to infer human Affect
Biologically Inspired approach
Functionally Designed approach
The scope of the thesis
Employing visual, auditory, tactile, and physiological signals to perform human-oriented perceptual tasks including; people detection, recognition, and tracking, gaze tracking, face detection and recognition, gesture recognition and body language understanding, speech and speaker recognition and natural language understanding

Figure 2.1 **Thesis scope and its research area**

The role of a perceptual system is central in many cognitive architectures as it is regarded the backbone upon which social robots acquire higher-level *social intelligence* skills. We will review the literature on plausibility of designing perceptual system for social robots in the light of understanding the perception process in human's brain from neuroscience, psychology, and psychophysics perspective.

The evolution of social robots can be traced back to the seminal studies in behaviouristic robotics published in the latter part of the last century. The pioneering research of Brook [23] and Arkin [24] led to the foundation of behavior robotics and in a parallel development, Allen Newell and John R. Anderson's studies led to birth of cognitive architectures [25]. Primarily inspired by biological species; the former focuses on design of robots that predominantly behave in accordance with the sense-act strategy. Hence, such robots do not require an explicit model of the environment. On the contrary, the latter is inspired by humans and aims to mimic the human intellectual infrastructure via computational models perceived to be close to the operation of the human brain.

The majority of robotics literature, however, is dominated by solutions for specific problems such as navigation, path planning, object recognition, etc., via probabilistic, model-based [26], [27], and soft computing methods [28]–[30]. Such skills are regarded as integral components of the social robots and mostly assumed already innate in the robot. However, robot navigation strategies in human social settings (human-aware robot navigation) requires additional skills such as respecting personal zone, respecting affordance space, and respecting etiquette for approaching people to name few [31], [32]. Designing and developing a human-aware navigation system is a crucial factor for the success of deploying social robots in household settings, offices, hospitals, etc. Most of the human-aware navigation architectures can be classified into the well-known robotics paradigm; reactive, deliberative, and hybrid architectures [33], [34]. Perception of the environment is a vital component in all these scenarios; particularly, human motion prediction which involves the capability of mental perception or what is known as theory of mind in psychology [35].

2.2. The role of HRI in social robotics

Human-Robot Interaction (HRI) is “a domain of study dedicated to understanding, designing, and evaluating robotic systems for use by or with humans " [36]. Hence, this field of study represents the heart of social robotics research area. Developing an efficient perceptual system plays a vital role in the success of a social robot in achieving its intended purpose in various social HRI scenarios [15], [37]. As social robots are expected to provide physical, mental, and/or social support to humans, the need for close spatial and temporal association between the two in all aspects of their interaction is important and is referred to as proximate interaction [36]. Similarly, in the context of peer-to-peer collaboration, the mixed-initiative interaction which often takes the form of dialogue interaction [38] is emphasized and is regarded as relevant research in social robotics as shown in Figure 2.2. In this type of human-robot interaction, the role of the robot is shifted from being an operator of something to imitate and/or act as a peer or mentor to a human. Thus, an essential element that leverages the performance of the social robots in achieving mixed-initiative interaction is the availability of an effective perceptual system capable of perceiving the state of the real-world environment and has the capacity to grasp the social and emotional cues (i.e. providing social situation assessment) in a natural and real-time fashion.

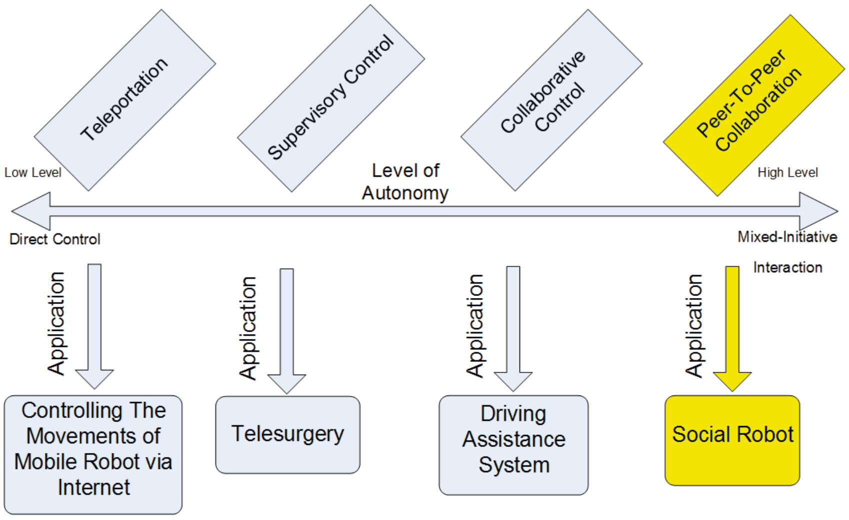


Figure 2.2 Levels of autonomy in different scenario of human-robot interaction

Dautenhahn [21] suggests the notion of a conceptual space of HRI and explains how various paradigms of social robots in literature are located within this space. The conceptual space is represented as a triangle where three HRI approaches namely; robot-centered HRI approach, human-centered approach, and robot cognition-centered approach represent the corner of the triangle and various social robots can be designed in accordance with the specific relationships among these approaches. Five social robot classes are defined and localized within this conceptual space of HRI including; socially evocative robots, socially situated robots, sociable robots, socially interactive robot, and socially intelligent robots as shown in figure 2.3. For example, socially intelligent robots, which are defined as robots that demonstrate aspects of human-style social intelligence, are designed based on deep model of human cognition and social competence [21] and are placed in the center of the triangle. These robots need to behave, communicate, interact, and look similar to humans as well as being equipped with explicit models of social cognition, interaction, and communication inspired by humans. Moreover, the way that humans perceive and respond to these robots is equally important, therefore, the three HRI approaches contribute to the development of socially intelligent robots. Even though significant progress in human-like appearance has been achieved, robots' sensory-motor control, cognition, communication, and interaction skills are still limited compared with the human level. Hence, building a socially intelligent robot that looks and behaves like a human is an ultimate and somewhat remote goal. Kismet, one of the early social robots developed at Massachusetts Institute of Technology by Breazeal [39], is placed at the corner of the robot-centered approach. This robot is designed as a creature that employs its interactions with humans to fulfill some of its needs (internal needs such as motivations, drives, emotions and external needs such as social needs) in order to pursue its own goals [20]. Socially interactive robots are compatible with human-centered approach where the human perspective is emphasized by designing robots that can fulfill their intended roles in a way that are acceptable and comfortable to targeted users. 'Users studies' in HRI, which fits the human-centered approach, can be used to facilitate designing social interactive robots by providing rough design guidelines that limit the design space of a specific robot for a particular task. Nevertheless, the exhaustive approaches of using HRI 'users studies' is not feasible due to high-dimensional characteristic of the design space for these robots [40]. The aforementioned definitions of social robots share one essential feature, that is, the ability of the robots to communicate, interact, and behave in social and in a natural way. Hence, social robots need to be endowed with social-emotional and social-cognitive skills to manifest human-like communicative, interactive, and behaviour competences as highlighted in figure 2.1.

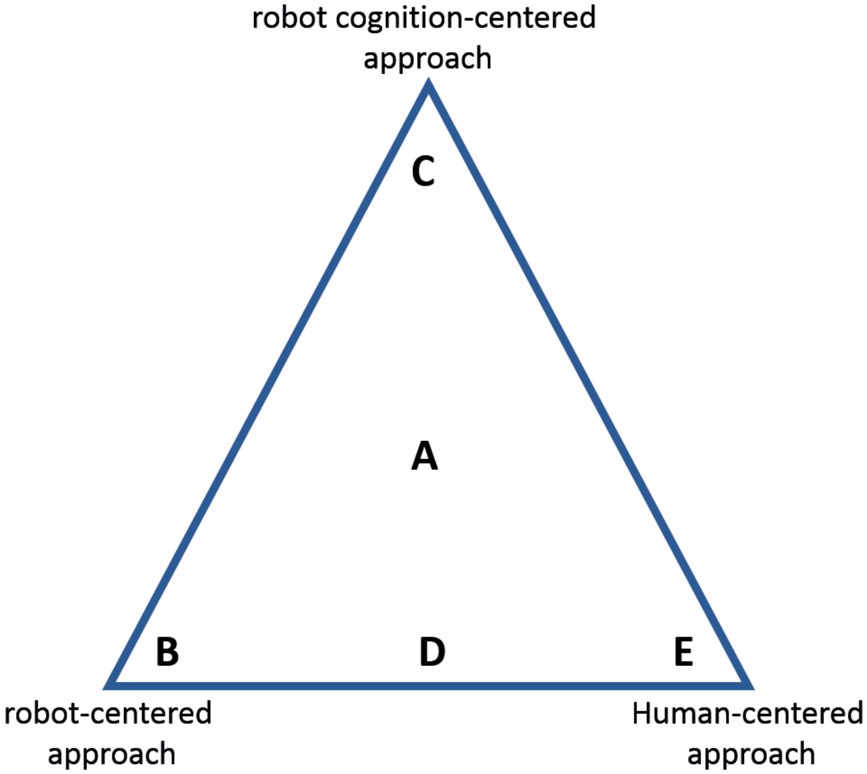


Figure 2.3 The conceptual space of HRI approaches. A, socially intelligent robots; B, sociable robots; C, socially interactive robots ; D, socially situated; E, socially evocative.

2.3. The role of affective computing in social robotics

Social-emotional skills are the ability of robots to detect, recognize, and interpret a person's affect, which is a combination of emotions, moods, attitudes, personality traits, in order to respond socially and predict person's future action during social HRI. Social-emotional skills include abilities that not only can identify a person's affective state but also possess explicit emotional models to communicate with their own internal states. Affective computing research area provides algorithms and computational models that can be integrated in social robots to manifest social-emotional skills. These algorithms and computational models employ visual, auditory, tactile, and physiological signals to infer the human affect. Two leading models have been adopted to characterize the human affect: categorical model and continuous model [41]. In categorical model, the perception of a person affective state is expressed as a multi-class classification problem where each class represents a finite number of affective states. The most common number of discrete affective states is six including; happiness, sadness, fear, disgust, surprise, and anger [42]. On contrast, continuous models define the affective state as a feature vector in multi-dimensional space. In this model, a person affect is conceptualized by localizing it in multi-dimensional space, such as valence and arousal [43], pleasure, arousal, and dominance [44], valence, arousal, and power [45]. Most of the continuous models are either two-dimensional or three-dimensional. A comprehensive survey on the most common modalities, feature vectors, and classification methods that could be utilized to detect and recognize a person affective state in social HRI was presented in [46]. Most of the research works in affective computing adopt machine learning techniques such as support vector machine, Gaussian mixture model, hidden Markov model, artificial neural network, and *k*-nearest neighborhood to perceive a person affective state. It is worth to highlight that recognizing a person affective state is needed to be integrated in cognitive architecture in order to demonstrate higher-level social-emotional skills. For example, a perceptual system for social robots is developed by Cominelli et al. [47], and it is integrated by the same research group as a key sub-system in a modular cognitive system referred to as the Social Emotional Artificial Intelligence (SEAI) [48]. Most of these perceptual systems adopted well-known machine learning algorithms and computational models that were developed originally as general-purpose or for other applications that do not fit the challenges and nature of social HRI.

2.4. The role of cognitive architecture in social robotics

The well-known cognitive architectures including ACT-R [25], CLARION [49], LIDA [50], ADAPT [51], and SOAR [52] have been suggested by cognitive and computer scientists to understand and replicate higher-level cognitive skills in human. Recently, a few number of cognitive architectures have been implemented on social robots [48], [53]–[56]. Most of these cognitive architectures simulate higher-level human cognitive skills by forming the mental abilities including; perception, attention, emotion, memory, learning, concept formation, and reasoning as mental modules and describing the interrelationships and communication pathways between these modules. Kismet is one of the social robots that has also been equipped with a cognitive system. It is developed by Cynthia Breazeal and it is called “the robot’s synthetic nervous system” (SNS) [53]. The system is designed as modules that dedicated to provide Kismet with the ability to exhibit animate features and to perceive social cues and human affect to allow the robot to be socially situated with people. Some cognitive architectures have also been developed based on the so-called Distributed Adaptive Control (DAC) theory and are implemented in various social robots including; iCub, Zeno [57], and Nao [58]. DAC theory suggests a neural architecture organized in two complementary structures (layers and columns) in order to replicate the modularity of the functions of brain areas and neural pathways for mimicking human cognitive skills. A modular cognitive architecture based on hybrid deliberative/reactive approach has also been reported as a Facial Automaton for Conveying Emotions(FACE) robot [55]. ACT-R/E is an embodied cognitive architecture that is adapted from the classical ACT-R architecture by adding one main condition on cognition. Cognition occurs within a physical body that perceives, interacts with, navigates in, and manipulates its environment [56]. In all of these cognitive architectures, the perception module is one of the main modules that is designed to fulfill the specific purpose for which these architectures have been conceived.

2.5. Human perception

Perusal through the relevant literature of perception methods for social HRI, it is interesting to note that most available methodologies do not take into consideration the fact that humans interact among themselves and with the environment through efficient processing of available information from multisensory modalities, and the integration of this information over the normal perception time [59], [60]. For example, in the person identification problem, humans use various sensor modalities and different aspects of measures to facilitate their perception and respond to stimuli within the norm of their interaction rate [61]. Facial expressions recognition, people detection, recognition, and tracking, gaze tracking, face detection and recognition, gesture recognition and body language understanding, speech and speaker recognition, and natural language understanding are considered among the main difficulties to be addressed by the perceptual system of a social robot to be able to mimic human-like interaction for a successful social human-robot interaction.

Research in social robotics is fundamentally interdisciplinary. Here, we advocate that the area can significantly benefit from the latest developments in neuroscience and psychology, computational cognition, and psychophysics. Social robots can interact with humans and the environment in a similar fashion to human interaction if they are embedded with a perceptual system that perceives the real-world as closely as humans do. To design these robots with similar perceptual capabilities as humans, not only should they be equipped with an analogous sensory system as humans, they also ought to mimic a closely related architecture to human's perception process [17]. Within the spirit of this principle, we are inspired by two interrelated systems in the human's brain: (1) the microstructure circuit of the human nervous system and its computational functions which are used to design a set of Dedicated Processing Units (DPUs) superior in processing specific feature vectors and (2) the macrostructure of the primate sensory cortex in order to emulate the binding mechanism that is used to integrate the outputs of these DPUs and to finalize the outcome of the perception process. Marr's Tri-level of analysis [62] (computational, algorithmic, and implementation levels) for information processing of vision system coincides with these two principles. On the one hand, the computational functions of the human nervous system reflect Marr's computational level where efficient technical algorithms are adopted to realize relevant cognitive function, and the underlying biological mechanisms of cognition are ignored. On the other hand, adhering to microstructure circuit of the human nervous system and macrostructure of the primate sensory cortex are consistent with the Marr's algorithmic level where the representations, mechanisms, and algorithms that need to implement the computation represent the core of this level of analysis.

Extensive review of the literature that provides plausible explanation of the human perception process and offers mechanisms for solving the binding problem was a crucial step in building the architecture reported in this thesis. Having said that, despite significant progress in understanding the human perceptual system, research in mind and perception process are still open and active research topics in many diverse fields including neuroscience, psychophysics, psychology, sociology, and cognitive science. The key findings from these research fields suggest possible models, architectures, and mechanisms of human perception process operation [3], [63]–[71]. These interpretations and mechanisms were used as basis of deriving the guidelines that have shaped the framework of a general-purpose human-oriented perceptual system.

Visual perception is considered a crucial element in the human perceptual system and plays a vital role in achieving many tasks including but not limited to person identification, facial expression recognition, and object recognition to name a few [72]–

[74]. The notion of hierarchical architecture of primate visual cortex has been established in the pioneering work of Hubel and Wiesel [62]. By studying the neural response to visual stimulus at different levels of hierarchical organization of a cat visual cortex (primary visual cortex V1 and extrastriate visual areas V2, V4, and IT), they found that most neurons in V1 responded strongly to bar-like stimuli at a particular orientation and position in the visual field. Also, they found that the so-called complex cells at higher level in the hierarchical architecture responded best to bar-like stimuli at a particular orientation regardless of their position in the visual field. Other studies discovered that neurons at higher level in the hierarchical organization of visual cortex responded strongly to complex shapes such as star-like shapes (IT area) [62] and faces at anterior inferotemporal cortex (AIT) area [62]. Although these studies found that neurons at higher levels are insensitive to some changes in the visual stimulus such as size, location, and contrast; other studies [75], [76] reported that the neurons at AIT area show view-dependent behaviours. The key finding of these studies inspired computer vision researchers to adhere to the main principle that the visual stimulus is processed via a feedforward hierarchical architecture whereas the complexity and invariance of the features that extracted form a visual stimulus are growing up going from early stage toward later stage in a hierarchical architecture.

One of the prominent models that was proposed by neuroscientists to explain and provide mechanism for the human perception process is the “convergence-zone”. In this model, the “convergence-zone” ensembles which are presumed to be located in the higher-order integrative cortical areas play a major role in integrating multiple aspects of the internal and external reality of perceiving and recalling experiences. In Damasio’s model [3] , a real-world entity, which has limited set of features such as color, texture, smell and taste, is encoded as a set of feature vectors. Each feature vector is composed of a number of elements that represent its associated characteristics. The central observation of Damasio’s model is that the real-world entities are perceived by a synergistic process that uses information extracted from various sensor modalities and integrate this information at a multi-level structure through a hierarchical manner. Each sensor modality provides one or more of the entity’s features and the hierarchical integration of this sensory information over time is the key factor that makes humans superior in perceiving and distinguishing the real-world objects in diverse settings and scenarios. For example, the vision modality provides features that characterize a real-world object such as color, shape, and size. In the human perception process, all the attributes, that belong to a certain object and available in sensor modalities, are encoded as a set of features vectors. These features vectors are fed to dedicated neural circuits to be processed. Then, the synergistic integration of the outputs of these dedicated neural networks, at different levels using different binding criteria, produces the final output of the perception process. The areas where the binding process is performed are called "convergence-zone" ensembles. The question raised by this model is how the "convergence-zone" ensembles binding the outputs of these neural networks to create a perception or experience about a certain object or event respectively. The Feature Integration Theory (FIT) – proposed by Terisman and Gelade, suggests a solution for the binding problem and attempts to answer this question [70] . FIT provides a mechanism for binding by location and shared features. In this theory, the spatial attention window provides access to the features in the associated location in various feature maps and facilitates the integration of information from these locations to a single object file for more analysis and identification. One of the key principles of FIT, which we have implemented in our model, is that objects features are "registered early, automatically, and in parallel across visual fields, while objects are identified separately" and at later stage in the perception process [70]. Another prominent solution of binding problem was formulated first by Milner, Grossberg , and von der Malsburg in [77][78], and [79] respectively, but Gray and Singer were the first who experimentally demonstrated the role of synchrony in the binding process [80]. They coined the term "Binding-by-synchrony" and suggested that the binding problem could be solved by the temporal synchrony of population of neurons. The neurons that encode the same object are distinguished from other neurons by synchronous firing. In other words, the matching frequency firing of population of neurons indicates that these neurons encode the same object, while other frequency firing highlight other objects. The proposed approach of developing human-oriented perceptual system employs the "convergence-zone" ensembles, FIT, and Binding-by-synchrony, in a synergistic fashion and in a complementary manner. The spatial window attention is used to register stimulus's features in parallel by feeding the feature vectors that represent the attributes of the attended stimulus to parallel distributed processing units, while the temporal binding is used at various hierarchical stages in order to integrate the outputs of these parallel distributed processing units which in turn create a perception of an attended object or an event.

The findings form neuroscience and psychophysics suggest that the formation of cell assemblies is controlled by the following principles: (1) population of neurons in a specific cell assembly must have similar receptive field properties, (2) each cell assembly maps one feature or quality of the attended stimulus, and (3) population of neurons in same cell assembly fire in temporal synchrony with each other. The first property can be realized by connecting each sensor modality to a dedicated receptive field system which in turn generates a set of feature vectors representing a set of qualities of attended stimulus. Feeding each of these feature vectors to its Dedicated Processing Units (DPU) satisfies the second principle. It should be noted that the last two principles can be manifested as hardwired property in the architecture of the human-oriented perceptual system where sensory systems are connected to their corresponding receptive field systems. The third principle, which states that the population of neurons in the same cell assembly fire in temporal synchrony with each other, can be engineered by equipping the social robot with the state-of-art sensors that are designed to provide information about the attended stimuli on the basis of event-driven computation [81].

Selection of the appropriate computational models of these DPUs is a key factor in accomplishing many tasks that social robots are proposed to achieve efficiently in social HRI perspective such as person identification, gesture recognition, facial expressions to name a few. The Marr's three level of analysis provides a framework for selecting appropriate computational models of these DPUs, particularly the computational level of analysis advocates that the practical progress in machine learning, speech and speaker recognition, and computer vision to solve perceptual tasks should contribute to understanding perception by identifying which feature set, classifier model, and training algorithm seem to work for a particular perceptual task.

In contrast to classical theories of sensory processing which consider the brain as a stimulus-driven mechanism, the current findings consider the human perception process as an active as well as a proactive process. Within this framework, the processing of stimuli is controlled by top-down influences that shape the outcome of the sensory processing by creating predication about the forthcoming sensory events and providing shortlisted candidates that used in searching of the pattern represent the attended stimulus. The effect of top-down influences is to alter or multiplex the function of neurons at the receptive field system according to the object or the task that is attended to [82]. In other words, the expectations, which are given as pre-knowledge or induced by the outcome of processing the feature vectors in the fastest processing routes, generate top-down influences by instructing receptive field system to create a set of appropriate feature vectors that are processed in hierarchical feedforward pathways in order to refine the attended object and complete the perceptual task at hand. This interpretation is supported by neuroscience and cognitive studies that consider perception as a constructive and active process [67][83]. They suggest that when human sensory systems such as vision, auditory, and tactile are excited by a real-world stimulus, the corresponding neural systems of these sensory systems produce mapping for the stimulus attributes that available in the modalities of these sensory systems. Since these patterns are mapped by population of neurons distributed across and within cortical hierarchy, the binding or perceptual grouping is accomplished by synchronization of neural firing among population of neurons that form the cell assembly or what we refer to it in our proposed model as dedicated processing units. Then, the integration of the outputs of these cell assemblies in parallel with the search for the best match of the attended pattern within the library of representations that stored in memory retrieves the full details of the attended stimulus.

2.6. Biological plausible computational models of human perception

Many researchers have adopted biologically inspired approaches to develop feature extraction methods, computational models, and architectures that employ data from sensory modalities in order to create a perception about an attended stimulus. Here, we present selective studies from computer vision and speech processing research domains that inspired by human sensory cortex, particularly, visual and auditory cortex.

2.6.1. Visual Modality

A huge part of research on computer vison has been inspired by the human visual process. Such systems are employed to develop perceptual systems for object recognition, visual segmentation, face recognition, gesture recognition, person recognition, and tracking, etc. Deep Convolutional Neural Networks (DCNN), which have demonstrated impressive results for object recognition using the ImageNet Large Scale Recognition Challenge dataset [84], are consistent with processing of information via visual cortex in two essential features: the hierarchical structure and increasing of neurons selectivity at the size of their receptive fields over different layers. However, data representation and solely feedforward connections in DCNN (i.e. the absence of recurrent and top-down connections) are not biologically plausible with visual cortex. HMAX and enhanced HMAX models [85], which are invariant object recognition systems and inspired from primate visual cortex, share two features with the visual cortex: the hierarchical structure and the biological plausible way to achieve selectivity and invariance in the hierarchical layers. HMAX model and its enhanced version show relatively high performance on invariant single object recognition, multi-class categorization, and complex scene understanding tasks [86], [87]. All the aforementioned models adopt mean firing strength for neural encoding of information. Recently, more biological plausible systems, which employ spike timing for neural encoding of information as well as hierarchical architecture, have been developed to solve the problems of object and face recognition [81]. Spiking deep convolutional neural networks, which share hierarchical architecture with HMAX and DCNN, however, these neural networks utilize spiking neuron models and use special learning algorithms (e.g. spike timing dependent plasticity) are different from those employed in DCNN. Machine learning methods (support vector machine, Gaussian mixture model, artificial neural networks, and k-nearest neighbors, etc.,) that are used to perform the aforementioned perceptual tasks can be adopted at the functional level of the perceptual system and at the same time be compatible with the Marr’s computational level of analysis. These systems assume predominantly a feedforward configuration with no feedback in their hierarchical architecture. However, extensive research studies emphasize that perception is an active process that incorporates lateral and feedback connections as well as top-down influences as essential elements for visual processing such as attentional focus, context awareness, expectations, and perceptual tasks [88].

Most of the above systems with cortex-like architecture have demonstrated impressive results in classification and recognition tasks, recently, DCNN have demonstrated superior performance in single object recognition in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [89]. However, it has been reported [90] that the GoogLeNet, the wining architecture of object recognition in the ILSVRC 2014, performance degraded dramatically when it was evaluated on stimulus representing drawings of everyday objects [91], on silhouette versions of these drawings [91], or and stimulus representing partially occluded shapes [90]. Moreover, recent studies demonstrated that CNNs learn class manifolds that are loose or not constrained in the same way as representations that developed by humans. For example, a yellow-and-black-striped image is recognized with a high probability as a school bus [92]. Finally, these CNN models still demonstrate low performance with localizing task (i.e. determine the location of the detected object in an image) or detecting all object in an image (approximately 75% and 44%, respectively, on the ILSVRC 2014 challenge) [84]. One main shortcoming of the current DCNNs is that the spatial dependencies between image regions (i.e. spatial structure and configural information) is not considered in the learning process and are lost at the higher stages in the hierarchical architecture [93], [94]. Many research studies emphasize the crucial role of configural information in categorization and recognition task such as scene categorization and face recognition [95]–[98]. Recently, the ability of Recurrent Neural Network (RNN) to encode contextual information among sequential data has been exploited to account for configural information processing of intermediate features of DCNN [93]. However, the performance of hybrid architecture of RNN and CNN still is modest in indoor and outdoor scene categorization tasks [94]. Most of recognition systems that adopt these architectures employ only one modality such as visual modality (e.g. face recognition, object recognition, etc.), auditory modality (e.g. speech recognition, speaker recognition, etc.), or tactile modality.

2.6.2. Auditory Modality

Auditory modality plays a central role in human communications. Human utilizes voice characteristics including glottal and vocal tract response, syllable stress, intonation patterns, speaking rate and rhythm, and lexicon to perceive what is being said and who is speaking seamlessly. Speech and speaker recognition research communities have adopted approaches that employ various voice characteristics to answer the aforementioned questions. The way that a computer processes speech information is inherently different from the underlying biological computation in the human auditory system. The state-of-art of speech and speaker recognition approaches have utilized statistically-based computational model (Gaussian mixture model, support vector machine, hidden Markov model) to process the extracted feature vectors from speech data. However, the development of the most common feature extraction methods such as Mel-Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction (PLP) coefficients, and Gammatone Frequency Cepstral Coefficients (GFCC), is inspired by the psychoacoustic and physiological results that revealed some insight about the underlying biological computation of human auditory system. The Mel scale, the Bark scale, and the ERB scale were adopted to design filter-banks that utilized in the feature extraction method of MFCC, PLP, and GFCC respectively. These scales were developed based on the results of clever psych-acoustical experiments ERB scale [99], Bark scale [100], and Mel scale [101]. The filter-banks designed by these scales are condensed at low frequency (below 1 kHz) and are lengthy-spaced at other frequency (above 1 KHz) which is consistent with physiological data of auditory processing. In addition, both MFCC and PLP feature extraction methods employ psycho-acoustically motivated amplitude compression method (the log scale for MFCC , and power-low compression For PLP) [102].

Recently, research groups at the University of Toronto, Microsoft Research , Google, and IBM Research have successfully utilized DNNs for acoustic modeling and particularly for developing automatic speech and speaker recognition system [103]. These DNN models demonstrate better performance over the conventional GMM-HMM approach when tested on five different large-vocabulary continuous speech recognition [103]. The shortcoming of DNNs to encode the temporal dynamic of sequential data as in the case of speech signal , which is inherently dynamic, suggest integrating recurrent neural network, which is a dynamic system, with deep architecture to encode the temporal dynamic of speech data and consequently improves the performance of automatic speech recognition system [104], [105]. A sophisticated bioinspired model, which incorporates long-short-term memory cells in the RNN architecture to increase the capability of RNN to model complex temporal dynamics, has demonstrated reasonable performance in various speech processing tasks [106][107]. Incorporating long-short-term memory cells in the RNN not only provides more biologically plausible model, but also demonstrate better ability to solve difficult tasks involving recognition of temporally extended patterns in noisy input sequences which are previously unsolvable by traditional RNN [108].

2.7. Conclusion Remarks

In the light of the above discussion, one can tentatively conclude that most of the perceptual algorithms may not be appropriate for perception within the context of social HRI. One of the significant challenges of perception in social HRI is that the reliability and the availability of each sensory modality varies according to HRI scenario, environmental conditions, or the attended perceptual task. For example, the vision modality is more reliable that auditory modality in recognizing a person in a noisy cocktail party. The asynchronous nature of processing information from sensory modalities, which can be due to variation in the processing time and the availability of these sensory modalities within a finite time window, is not considered in many reported multi-modal recognition systems. However, the later parameter is very important in social HRI in order to facilitate the response of the social robot and to respond in real-time fashion even when one or more modality fail, and/or the desired reliability of the perception outcome is not satisfied. Finally, the lateral and feedback connections play crucial roles in manifesting top-down influences that represent very important aspect of perceptual and cognitive skill and should be incorporated in designing a perceptual system for social robot. We believe that the multimodal perception and asynchrony nature of processing information from sensory modalities as well as the importance of time dimensionality and top-down influences to facilitate social HRI are essential for developing perceptual system for social robots. This will likely fulfil some features of socially intelligent robots as more and more robots will enter the humans living environments including; homes, offices, and hospitals, etc.

Chapter 3. Summary of Contributions

The main contributions of this Ph.D. research project are presented in three peer-reviewed journal papers two of which have already been published and the third one is under review at the time of writing this thesis. They are reproduced in Appendix A-C. In this chapter, a summary of contribution of each paper is reported.

3.1. Architecture of a general-purpose human-inspired perceptual system for social robots

The remarkable capacity of human to perceive its environment in diverse situations based on incomplete and at times vague information has been linked to an underlying sophisticated perceptual system. As social robots are expected to function predominantly in human environments and interact with humans; it is most likely and desirable that these robots not only need to be equipped with similar sensory system as humans; they also ought to mimic a closely related architecture to human's perception process. The key contributions of this work [109] are as follows:

- A novel perceptual system is reported as a large-scale end-to-end social robotic perceptual architecture along with design guidelines to facilitate applying the proposed architecture to many on-going problems in social robotics. The key sub-system of the perceptual system and interrelationship among them as well as a one-to-one analogy of these sub-systems to the biological system of human sensory cortex are introduced in section III (please refer to Figure 1 for overview of the system).

- The research proposes an elegant way to reduce the error bias and variance by introducing built-in feature in the proposed architecture, that is, various kind of information are extracted from each sensor modality and are fed to their associated DPU for further processing (please refer to Figure 2 and section III-A).

- We also present a plausible mechanism for processing multi-modal sensory information by introducing top-down influences feature in the architecture as two kinds of connections (lateral and feedback connections) in order to reduce the computational cost of the perceptual system and facilitate real-time response in social HRI scenarios (please refer to Figure 3 & 4, section III-B, and section III-C).

- Two algorithmic realization of integrating information from multisensory streams via temporal-binding with fading memory are suggested in this work; temporal-binding with fading memory via liquid state machine and temporal-binding with fading memory via leaky integrate-and-fire neuron (please refer to Figure 5 & 6 and section III-D).

For complete and detailed information about the system, the reader is referred to Appendix A or [109]

3.2. Multi-modal person recognition system for social robots

Recognizing people in diverse social settings, an indispensable attribute that is often taken for granted, yet playing a central role in our social interactions. In the context of social robotics, it is very much desired that social robots, like humans, seamlessly identify familiar persons in their social circles without any intrusive biometric verification procedure. The main contributions of this work [110] can be summarized as follows:

- In this work [110], a solution to the problem of person recognition in social settings via a customized version of the multi-modal perceptual system is presented (please refer to Figure 2 and section 3). The system incorporates multimodal biometrics features that are non-invasive, are not affected by changes in appearance (i.e., outfit change), and works within the range of human social interaction rate (human response time) to address the challenges associated with problem of person recognition in social settings. In particular, the system utilizes configural- and appearance-based information in vision modality as well as short-term spectral information in voice modality to recognize a person in various social human-robot interaction scenarios (please refer to section 3.2.1 and section 3.2.2).

- As most reported multi-biometric person recognition algorithms require the presence of all modalities to finalize the person recognition task which make them not appropriate for social human-robot interactions. However, the proposed system relaxes this constraint by employing spiking neurons to integrate the information from various sensor modalities in order to overcome the challenge of asynchrony nature of processing information in social HRI scenarios (please refer to Figure 8 and section 3.4).

- A hybrid multimodal database is formed by integration of the ubiquitous FERET, RGB-D, and TIDIGITS datasets for face recognition, person recognition, and speaker recognition, respectively (please refer to section 4.1).

- This multimodal dataset is employed for testing the algorithm and assessing its merits against related methods. Within the context of the social robotics, the results suggest the superiority of the proposed method over other person recognition algorithms (please refer to section 4 and section 5).

For complete and detailed information about these studies, the reader is referred to Appendix B or [110].

3.3. Experimental realization of the proposed perceptual system and the effect of incorporating top-down influences on its performance

A multi-modal human-oriented perceptual system is realized in an in-house designed social robotic platform to demonstrate the properties and to evaluate the performance of the proposed perceptual system on real-time social HRI scenarios [109], The highlights of this work are as follows:

- A realization of social robotics platform is presented. The social robot is equipped with a RGB-D sensor (Kinect sensor), RGB camera (USB3 Flea3 machine vision camera), and Directional microphone (Rode VideoMic Pro) to represent visual and auditory modality (please refer to Figure 9 and section IV-A).

- A customization of the perceptual system to solve the problem of person recognition in social settings is introduced. The system employs the information from robot's visual modality (RGB-D sensor and RGB camera) and auditory modality (Directional microphone) to identify a person in divers social HRI scenarios (please refer to Figure 8 and section IV-B).

- An experimental realization of top-down influences to reduce the computational cost and to improve the recognition accuracy of the proposed perceptual system in person identification task was also demonstrated (please refer to section IV-B-2). Quick response (QR) codes were adopted as a sensible and straightforward solution to infer crucial information about the environment. The cues from QR supplies complementary features to simplify the person recognition task. QR codes were assigned with a string to identify the associated laboratory and the individuals who normally work therein (please refer to Figure 10). Once the QR codes are detected and decoded by the robot's vision system, the social robot gets prior knowledge on the specific group of individuals expected to be in that area. Hence, limited number of spiking neurons are biased in order to reduce search space and to lessen computational burden. These spiking neurons receive additional inputs (i.e. the biometric features that provided by other sensory modalities and processed by their associated DPUs) and compete among each other to represent the best candidate of the attended subject (please refer to section IV-B-3).

- The performance of the multi-modal social robot was evaluated in different scenarios whereby one or more modalities are not available (please refer to section IV-C and section IV-D). To demonstrate the impact of incorporating top-down influences on the performance of the system, the system was evaluated in two cases; in the first case, the QR codes were not posted on doors (i.e. incorporating top-down influences), while in the second case, the QR codes were posted on doors ((i.e. not incorporating top-down influences). The results show that incorporation QR codes as top-down influences improve the recognition rate and facilitate real-time response (please refer to Figures 15 & 17).

For complete and detailed information about this work, the reader is referred to Appendix A or [109].

3.4. A two-stage classifier speaker recognition system based on prosodic and spectral features via fuzzy inference fusion

In the absence of a unique robust speaker identification system that demonstrates superior performance for applications where the system is expected to perform in challenging scenarios such as different types of environmental noise, at different levels of environmental noise (wide range of signal-to-noise ratio), and with only access to short utterances (at test time), the plausible contention is to integrate the advantages of using multi-feature speaker recognition system with multi-classifier speaker recognition system. The contributions of this work [111] are as follows:

- A novel architecture for speaker recognition system was developed in the light of the design guidelines that proposed to develop perceptual system for social robots. Utilizing different kinds of information within auditory modality and processing them via specialized computational models are manifested in the proposed architecture as exploiting multi-feature system (prosodic and short-term spectral features) and multi-classifier system (discriminative and generative models), respectively (please refer to Figure 1 and sections 3.1 & 3.2).

- The top-down influence feature of the perceptual system is employed as a means of adopting prosodic features to cluster the speaker population into two classes which in turn are used to build strong coupling between speaker-dependent model and universal background model and to reduce the population size of speakers (please refer to Figure 1 and section 3).

- The system employs on-fly fuzzy fusion system to combine the outputs of multi-feature multi-classifier systems in order to improve the performance of the system in challenging scenarios such as noisy environments and at times when only short utterances are available (please refer to section 3.3).

- The fuzzy inference system employs length of utterance and signal-to-noise ratio to compute the weights of the base classifiers relying upon IF-THEN rules that set by expert (i.e. IF-THEN rules are constructed by studying the performance of the base classifiers at different combinations of the aforementioned challenging conditions).

- Experimental evaluations based on TIDGITS database suggest that the proposed architecture is promising and can improve the recognition rate of the system in the aforementioned challenging conditions, particularly, where the signal-to-noise ratio is very low (very noisy environment) and the length of utterance is short (please refer to section 4).

For complete and detailed information about this work, the reader is referred to Appendix C or [111].

Chapter 4.
Conclusions and Future Work

4.1. Conclusions

A school of thought among scientists and philosophers warns us of a future dystopia in which humans are ruled by intelligent and senseless robots. The majority of robotics research, nonetheless, does not subscribe to such a bleak insinuation and instead envisages a world where autonomous and intelligent robots are assimilated into the human world and are part of the solution to alleviate man-made problems. In that sense, the emerging area of social robotics is primarily concerned with humanizing the robots. There are two facets in this intricate process: (1) designing autonomous and intelligent robots; (2) making robots that look like or act like humans and display human-like emotions. The former is concerned with engineering robots that have learning, decision-making capacities, and navigate seamlessly whereas the latter pursues the question of how human characteristics such as emotions and cognitions, seemingly not mechanical, can be simulated and embedded in mechanical machines. Together the two approaches are expected to evolve into design of robots that are able to pass the Turing Test [112]. There is abundance of research on autonomous and intelligent robots and impressive results are reported in both domains. The latter area, however, is gaining more attention particularly in the last decade. The inspirations and the benchmark for both, of course, are humans and bio-systems. In this context, we believe that neuroscience, psychophysics, computational cognition, and psychology contribute enormously as they have the key to above queries.

Human-like perceptual system is a key sub-system that social robots need to be endowed with to acquire high-level social intelligence skills. This thesis contributes to the domain of social robotics with a novel design framework towards human-like perceptual system for social robots. In contras to the reported methodologies that address perception as solely a pattern recognition problem without taking into account the social implications of the perceptual tasks, the proposed approach emphasizes the social perspective of the perception and incorporates it in the design process. The research presented in this thesis investigates how an interdisciplinary approach, which synthesizes the findings from psychology, psychophysics, and neuroscience about human perception in a design framework for developing human-oriented perceptual system, could offer elegant solutions to the main challenges of perception in social settings. The main contribution of the thesis includes but not limited to an end-to-end social robotics platform and a set of design guidelines and a general-purpose human-oriented social robotic perceptual architecture. The design guidelines and the architecture serve as platform to facilitate addressing many on-going perceptual tasks that needed to realize socially interactive robots. Additionally, this platform was adopted to address the problem of person recognition in social settings and the problem of speaker identification in noisy environment and with only access to short utterances at test time.

The thesis began with proposing an architecture for a general-purpose perceptual system for social robots. The design framework utilizes the macrostructure of the primate sensory cortex and the microstructure circuit of the human nervous system to realize a modest engineering replica analogous to multisensory information processing in human sensory cortex. The findings in neuroscience, psychology, and psychophysics about human perception process are incorporated in the design guidelines that shape the perceptual architecture for numerous perceptual tasks in the context of social HRI. Employing spiking neurons and top-down influences as hardwired properties in the proposed perceptual architecture offers some unique advantages for addressing the special challenges raised by the perception in social settings. The challenge of asynchronous nature of processing information in sensory modalities has been addressed by employing integration of information via spiking neurons that does not require the simultaneous presence of all considered cues of an attended stimulus and can facilitates a decision by any modality that is rich in information and first becomes available. Utilizing top-down influences to filter and condense the flood of data streams form sensory modalities as flexible representative map, which can be customized for the attended task by processing additional cues, facilities real-time response and reduce computational cost of the proposed system.

A customized version of the general-purpose perceptual architecture was configured to address the problem of person recognition in social settings. The system is multi-modal, non-invasive, and does not require that all input stimuli are simultaneously available. In addition, the proposed system has the ability to adapt to real-world scenarios of social HRI by adjusting the threshold value which compromises between the reliability of the perception outcome and the time required to finalize the perception process.

Extensive simulations and comparative studies to evaluate the performance of the proposed system on person recognition at different social HRI scenarios, suggest remarkable advantages over related methods for person recognition in social settings.

The above design guidelines have also been utilized to develop a robust speaker identification system. Employing low-dimensional prosodics feature vector to cluster the population of speakers into two groups (male and female) is consistent with notion of top-down influences. This feature reduces the population size and builds a strong coupling between speaker-dependent model and the UBM. Moreover, Integrating the advantages of using multi-feature (two types of short-term spectral features and prosodics features) speaker recognition system with multi-classifier (support vector machine, Gaussian mixture model) speaker recognition system in an elegant speaker identification architecture is consistent with the above design guidelines. The speaker recognition system has the following characteristics: 1) real-world stimulus is composed of a set of features that vary in their relative salience on the perception outcome and complement each other; 2) The set of features vectors that represents the attended stimulus are processed by DPU's. These DPU's have dedicated architectures, use different learning methods, and superior in extracting a specific type of information available in these feature vectors. The system demonstrates superior performance at different challenging conditions particularly at low signal-to-noise ratio (SNR) and short utterance.

4.2. Future Work

Due to the universality feature for the proposed perceptual system, there are several possibilities for further investigation and adaptation of the proposed perceptual system. Some of the paramount research directions that are noteworthy to investigate within the scope of this work are:

- Utilizing the proposed framework for developing multi-modal human affect recognition system for social robots. The system may employ numerous modalities such as facial expressions, body language, voice, and physiological signals to infer person's affective state in more accurate and robust way.
- Utilizing the proposed framework for developing multi-modal place recognition system for social robots. The data stream from Laser-rangefinders, infrared thermal camera, and RGB camera can be utilized within the proposed architecture to develop robust place recognition system.

- Designing multi-modal object recognition system that utilizes the stream of data from tactile sensor, depth sensor, and RGB camera to extract discriminant features such as texture information, configural information, and appearance information. Then, Integration of this information within the proposed architecture generates a comprehensive map for the attended object and consequently robust object recognition system.

- Deep convolutional neural network (DCNNs) have recently demonstrated excellent performance as a robust feature extractor for several perceptual tasks including scene categorization, object recognition, speech recognition [113], [114]. The ability of DCNNs to learn generic features that are transferrable to a variety of related tasks within the same modality, push in the direction of replacing appearance-based hand-crafted features with learned features that available at different intermediate levels of DCNN hierarchical architecture. However, the configural information is not encoded well in the current DCNNs. Hence, more research is needed to find a better method for configural-based generic features and integrate it within our proposed architecture.

4.3. Epilogue

At the onset of my research, I had so many questions and knew little outside engineering. I had to step out of my comfort zone, I struggled through the labyrinth of ideas, got lost here and there, and in the process, I leaned a bit about the inner working of the amazing human brain. Perception, of course, is only one part of the decision making. As such, designing a social robot that seamlessly communicates with human as we do among ourselves could be a mirage! I have a better understanding but have many open questions! The burning question, though, is: can a social robot perception system be as elaborate and flexible as humans to be described as a "*controlled hallucination*"?

References

[1] A. Treisman, “Properties, parts, and objects.,” in *Handbook of perception and human performance, Vol. 2: Cognitive processes and performance.*, Oxford, England: John Wiley & Sons, 1986, pp. 1–70.

[2] Stanford University. and Center for the Study of Language and Information (U.S.), “Stanford encyclopedia of philosophy.” Stanford University, 2011.

[3] A. R. Damasio, “The brain binds entities and events by multiregional activation from convergence zones,” *Neural Comput.*, vol. 1, no. 1, pp. 123–132, 1989.

[4] C.-A. Smarr, T. L. Mitzner, J. M. Beer, A. Prakash, T. L. Chen, C. C. Kemp, and W. a. Rogers, “Domestic Robots for Older Adults: Attitudes, Preferences, and Potential,” *Int. J. Soc. Robot.*, vol. 6, no. 2, pp. 229–247, Dec. 2013.

[5] J. Broekens, M. Heerink, and H. Rosendal, “Assistive social robots in elderly care: a review,” *Gerontechnology*, vol. 8, no. 2, Jun. 2009.

[6] A. Tapus, M. Mataric, and B. Scassellati, “Socially assistive robotics: The grand challenges in helping humans through social interaction,” *IEEE Robot. Autom. Mag.*, vol. 14, no. 1, pp. 35–42, 2007.

[7] K. C. Welch, U. Lahiri, Z. Warren, and N. Sarkar, “An Approach to the Design of Socially Acceptable Robots for Children with Autism Spectrum Disorders,” *Int. J. Soc. Robot.*, vol. 2, no. 4, pp. 391–403, Jul. 2010.

[8] T. L. Mitzner, T. L. Chen, C. C. Kemp, and W. a Rogers, “Identifying the Potential for Robotics to Assist Older Adults in Different Living Environments.,” *Int. J. Soc. Robot.*, vol. 6, no. 2, pp. 213–227, Apr. 2014.

[9] C. L. Lisetti, S. M. Brown, K. Alvarez, and A. H. Marpaung, “A Social Informatics Approach to Human-Robot InteractionWith a Service Social Robot,” *IEEE Trans. Syst. Man Cybern. Part C (Applications Rev.*, vol. 34, no. 2, pp. 195–209, 2004.

[10] V. Gonzalez-Pacheco, A. Ramey, F. Alonso-Martin, a. Castro-Gonzalez, and M. a.

Salichs, “Maggie: A Social Robot as a Gaming Platform,” *Int. J. Soc. Robot.*, vol. 3, no. 4, pp. 371–381, Sep. 2011.

[11] T. Kanda, M. Shiomi, Z. Miyashita, H. Ishiguro, and N. Hagita, “A communication robot in a shopping mall,” *IEEE Trans. Robot.*, vol. 26, no. 5, pp. 897–913, 2010.

[12] J. Pineau, M. Montemerlo, M. Pollack, N. Roy, and S. Thrun, “Towards robotic assistants in nursing homes: Challenges and results,” *Rob. Auton. Syst.*, vol. 42, no. 3–4, pp. 271–281, 2003.

[13] H. Gross, C. Schroeter, S. Mueller, M. Volkhardt, E. Einhorn, A. Bley, T. Langner, M. Merten, C. Huijnen, H. van den Heuvel, and others, “Further progress towards a home robot companion for people with mild cognitive impairment,” 2012, pp. 637–644.

[14] A. Niculescu, B. van Dijk, A. Nijholt, D. K. Limbu, S. L. See, and A. H. Y. Wong, “Socializing with Olivia, the Youngest Robot Receptionist Outside the Lab BT,” in *Social Robotics. ICSR 2010. Lecture Notes in Computer Science*, 2010, pp. 50–62.

[15] H. Yan, M. H. Ang, and A. N. Poo, “A Survey on Perception Methods for Human–Robot Interaction in Social Robots,” *Int. J. Soc. Robot.*, vol. 6, no. 1, pp. 85–119, Jul. 2013.

[16] T. Shiwa, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, “How Quickly Should a Communication Robot Respond? Delaying Strategies and Habituation Effects,” *Int. J. Soc. Robot.*, vol. 1, no. 2, pp. 141–155, Feb. 2009.

[17] T. Fong, I. Nourbakhsh, and K. Dautenhahn, “A survey of socially interactive robots,” *Rob. Auton. Syst.*, vol. 42, no. 3–4, pp. 143–166, Mar. 2003.

[18] C. Bartneck and J. Forlizzi, “A design-centred framework for social human-robot interaction,” in *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759)*, 2004, pp. 591–594.

[19] S. del Moral, D. Pardo, and C. Angulo, “Social Robot Paradigms: An Overview,” in *Bio-Inspired Systems: Computational and Ambient Intelligence*, 2009, pp. 773–780.

[20] C. Breazeal, “Toward sociable robots,” *Rob. Auton. Syst.*, vol. 42, no. 3, pp. 167–175, 2003.

[21] K. Dautenhahn, “Socially intelligent robots: Dimensions of human-robot interaction,” *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 362, no. 1480, pp. 679–704, 2007.

[22] J. Hirth, N. Schmitz, and K. Berns, “Towards Social Robots: Designing an Emotion-Based Architecture,” *Int. J. Soc. Robot.*, vol. 3, no. 3, pp. 273–290, Jan. 2011.

[23] R. Brooks, “A robust layered control system for a mobile robot,” *IEEE J. Robot. Autom.*, vol. 2, no. 1, pp. 14–23, 1986.

[24] R. C. Arkin and I. NetLibrary, *Behavior-based robotics*. Cambridge, Mass: MIT Press, 1998.

[25] J. R. Anderson, *How Can the Human Mind Occur in the Physical Universe?* New York: Oxford University Press, 2007.

[26] D. F. Sebastian Thrun, Wolfram Burgard, *Probabilistic robotics*. Cambridge, Mass. : MIT Press, 2005.

[27] J. F. Ferreira and J. Miranda Dias, *Probabilistic Approaches to Robotic Perception*, vol. 91. Cham: Springer International Publishing, Cham, 2014.

[28] C. Zhou, D. Maravall, and D. Ruan, *Autonomous robotic systems : soft computing and hard computing : methodologies and applications*. Heidelberg ; New York : Physica-Verlag, 2003.

[29] M.-R. Akbarzadeh-T, K. Kumbla, E. Tunstel, and M. Jamshidi, “Soft computing for autonomous robotic systems,” *Comput. Electr. Eng.*, vol. 26, no. 1, pp. 5–32, 2000.

[30] A. Saffiotti, “The uses of fuzzy logic in autonomous robot navigation,” *Soft Comput.*, vol. 1, no. 4, pp. 180–197, 1997.

[31] J. Rios-Martinez, A. Spalanzani, and C. Laugier, “From Proxemics Theory to Socially-Aware Navigation: A Survey,” *Int. J. Soc. Robot.*, vol. 7, no. 2, pp. 137–153, 2015.

[32] G. Ferrer, A. Zulueta, F. Cotarelo, and A. Sanfeliu, “Robot social-aware navigation framework to accompany people walking side-by-side,” *Auton. Robots*, vol. 41, no. 4, pp. 775–793, 2017.

[33] T. Kruse, A. K. Pandey, R. Alami, and A. Kirsch, “Human-aware robot navigation: A survey,” *Rob. Auton. Syst.*, vol. 61, no. 12, pp. 1726–1743, 2013.

[34] K. Charalampous, I. Kostavelis, and A. Gasteratos, “Recent trends in social aware robot navigation: A survey,” *Rob. Auton. Syst.*, vol. 93, pp. 85–104, 2017.

[35] B. Scassellati, “Theory of Mind for a Humanoid Robot,” *Auton. Robots*, vol. 12, no. 1, pp. 13–24, 2002.

[36] M. a. Goodrich and A. C. Schultz, “Human-Robot Interaction: A Survey,” *Found. Trends® Human-Computer Interact.*, vol. 1, no. 3, pp. 203–275, 2007.

[37] Y. Benezeth, B. Emile, H. Laurent, and C. Rosenberger, “Vision-Based System for Human Detection and Tracking in Indoor Environment,” *Int. J. Soc. Robot.*, vol. 2, no. 1, pp. 41–52, Dec. 2009.

[38] S. Jiang and R. C. Arkin, “Mixed-Initiative Human-Robot Interaction: Definition, Taxonomy, and Survey,” in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 2015, pp. 954–961.

[39] C. L. Breazeal and C. L. Breazeal, *Designing sociable robots*. Cambridge, Massachusetts : MIT Press, 2002.

[40] K. Dautenhahn, “Human-Robot Interaction,” in *The Encyclopedia of Human Computer Interaction*, 2nd ed., Soegaard; Mads and Dam; Rikke Friis, Ed. Aarhus, Denmark: The Interaction Design Foundation.

[41] A. Martinez and S. Du, “A Model of the Perception of Facial Expressions of Emotion by Humans: Research Overview and Perspectives,” *J. Mach. Learn. Res.*, vol. 13, p. 1589, 2012.

[42] P. Ekman, *Emotion in the human face*, 2nd ed. Cambridge Cambridgeshire, 1982.

[43] J. A. Russell, “A circumplex model of affect,” *J. Pers. Soc. Psychol.*, vol. 39, no. 6,

pp. 1161–1178, 1980.

[44] A. Mehrabian, “Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament,” *Curr. Psychol.*, vol. 14, no. 4, pp. 261–292, 1996.

[45] R. Plutchik and H. R. Conte, *Circumplex models of personality and emotions.* American Psychological Association, 1997.

[46] D. McColl, A. Hong, N. Hatakeyama, G. Nejat, and B. Benhabib, “A Survey of Autonomous Human Affect Detection Methods for Social Robots Engaged in Natural HRI,” *J. Intell. Robot. Syst. Theory Appl.*, vol. 82, no. 1, pp. 101–133, 2016.

[47] L. Cominelli, N. Carbonaro, D. Mazzei, R. Garofalo, A. Tognetti, and D. De Rossi, “A Multimodal Perception Framework for Users Emotional State Assessment in Social Robotics,” *Future Internet* , vol. 9, no. 3. 2017.

[48] L. Cominelli, D. Mazzei, and D. E. De Rossi, “SEAI: Social Emotional Artificial Intelligence Based on Damasio’s Theory of Mind ,” *Frontiers in Robotics and AI* , vol. 5. p. 6, 2018.

[49] R. Sun, “The CLARION Cognitive Architecture: Extending Cognitive Modeling to Social Simulation,” in *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation*, R. Sun, Ed. Cambridge: Cambridge University Press, 2005, pp. 79–100.

[50] S. Franklin and F. G. Patterson Jr, “The LIDA architecture: Adding new modes of learning to an intelligent, autonomous, software agent,” in *IDPT-2006 Proceedings (Integrated Design and Process Technology)*, 2006.

[51] D. P. Benjamin, D. M. Lyons, and D. W. Lonsdale, “ADAPT: A Cognitive Architecture for Robotics.,” in *the International Conference of Cognitive Modeling*, 2004, pp. 337–338.

[52] J. E. Laird, A. Newell, and P. S. Rosenbloom, “SOAR: An architecture for general intelligence,” *Artif. Intell.*, vol. 33, no. 1, pp. 1–64, 1987.

[53] C. Breazeal and B. Scassellati, “How to build robots that make friends and influence

people," in *Proceedings 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human and Environment Friendly Robots with High Intelligence and Emotional Quotients (Cat. No.99CH36289)*, 1999, vol. 2, pp. 858–863 vol.2.

[54] P. F. M. J. Verschure, "Distributed Adaptive Control: A theory of the Mind, Brain, Body Nexus," *Biol. Inspired Cogn. Archit.*, vol. 1, pp. 55–72, 2012.

[55] N. Lazzeri, D. Mazzei, L. Cominelli, A. Cisternino, and D. De Rossi, "Designing the Mind of a Social Robot," *Appl. Sci.*, vol. 8, no. 2, p. 302, 2018.

[56] G. Trafton, L. Hiatt, A. Harrison, F. Tanborello, S. Khemlani, and A. Schultz, "ACT-R/E: An Embodied Cognitive Architecture for Human-Robot Interaction," *J. Human-Robot Interact.*, vol. 2, no. 1, pp. 30–55, 2013.

[57] V. Vouloutsi, M. Blancas, R. Zucca, P. Omedas, D. Reidsma, D. Davison, V. Charisi, F. Wijnen, J. van der Meij, V. Evers, D. Cameron, S. Fernando, R. Moore, T. Prescott, D. Mazzei, M. Pieroni, L. Cominelli, R. Garofalo, D. De Rossi, and P. F. M. J. Verschure, "Towards a Synthetic Tutor Assistant: The EASEL Project and its Architecture," in *Conference on Biomimetic and Biohybrid Systems*, 2016, pp. 353–364.

[58] S. Fernando, E. C. Collins, A. Duff, R. K. Moore, P. F. M. J. Verschure, and T. J. Prescott, "Optimising Robot Personalities for Symbiotic Interaction," in *Conference on Biomimetic and Biohybrid Systems*, 2014, pp. 392–395.

[59] D. Nikolić, S. Häusler, W. Singer, and W. Maass, "Distributed fading memory for stimulus properties in the primary visual cortex.," *PLoS Biol.*, vol. 7, no. 12, p. e1000260, Dec. 2009.

[60] C. Spence and S. Squire, "Multisensory Integration: Maintaining the Perception of Synchrony," *Curr. Biol.*, vol. 13, no. 13, pp. R519–R521, Jul. 2003.

[61] A. Rice, P. J. Phillips, V. Natu, X. An, and A. J. O'Toole, "Unaware person recognition from the body when face identification fails.," *Psychol. Sci.*, vol. 24, no. 11, pp. 2235–43, Nov. 2013.

[62] D. Marr and D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. The MIT Press, 2010.

[63] M. Riesenhuber and T. Poggio, “Neural mechanisms of object recognition.,” *Curr. Opin. Neurobiol.*, vol. 12, no. 2, pp. 162–8, Apr. 2002.

[64] C. von der Malsburg, “Binding in models of perception and brain function.,” *Curr. Opin. Neurobiol.*, vol. 5, no. 4, pp. 520–6, Aug. 1995.

[65] a Treisman, “Solutions to the binding problem: progress through controversy and convergence.,” *Neuron*, vol. 24, no. 1, pp. 105–10, 111–25, Sep. 1999.

[66] A. Treisman, “The binding problem,” *Curr. Opin. Neurobiol.*, vol. 6, no. 2, pp. 171–178, 1996.

[67] W. Singer, “Neuronal synchrony: a versatile code for the definition of relations?,” *Neuron*, vol. 24, no. 1, pp. 49–65, Jun. 1999.

[68] A. R. Damasio, “Time-locked multiregional retroactivation: a systems-level proposal for the neural substrates of recall and recognition.,” *Cognition*, vol. 33, no. 1–2, pp. 25–62, Nov. 1989.

[69] N. Cowan, “The magical number 4 in short-term memory: a reconsideration of mental storage capacity.,” *Behav. Brain Sci.*, vol. 24, no. 1, pp. 87-114; discussion 114–85, Feb. 2001.

[70] A. M. Treisman and G. Gelade, “A feature-integration theory of attention,” *Cogn. Psychol.*, vol. 12, no. 1, pp. 97–136, Jan. 1980.

[71] V. Goffaux, B. Hault, C. Michel, Q. C. Vuong, and B. Rossion, “The respective role of low and high spatial frequencies in supporting configural and featural processing of faces,” *Perception*, vol. 34, no. 1, pp. 77–86, 2005.

[72] S. Gong, M. Cristani, S. Yan, and C. C. Loy, *Person Re-Identification*. London: Springer London, 2014.

[73] R. Chellappa, P. Sinha, and P. J. Phillips, “Face recognition by computers and humans,” *Computer (Long. Beach. Calif).*, vol. 43, no. 2, pp. 46–55, Jun. 2010.

[74] B. Fasel and J. Luettin, “Automatic facial expression analysis: a survey,” *Pattern Recognit.*, vol. 36, no. 1, pp. 259–275, Jan. 2003.

[75] H. H. BГjlthoff and S. Edelman, “Psychophysical support for a two-dimensional view interpolation theory of object recognition,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 89, no. 1, p. 60, 1992.

[76] N. K. Logothetis, J. Pauls, H. H. BГjlthoff, and T. Poggio, “View-dependent object recognition by monkeys,” *Curr. Biol.*, vol. 4, no. 5, pp. 401–414, 1994.

[77] P. M. Milner, “A model for visual shape recognition.,” *Psychol. Rev.*, vol. 81, no. 6, pp. 521–535, Nov. 1974.

[78] S. Grossberg, “Adaptive pattern classification and universal recoding: II. Feedback, expectation, olfaction, illusions,” *Biol. Cybern.*, vol. 23, no. 4, pp. 187–202, 1976.

[79] C. Von Der Malsburg, “The correlation theory of brain function,” in *Models of neural networks*, Springer, 1994, pp. 95–119.

[80] C. M. Gray and W. Singer, “Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex,” *Proc. Natl. Acad. Sci.* , vol. 86, no. 5, pp. 1698–1702, Mar. 1989.

[81] A. Delorme and S. J. Thorpe, “SpikeNET: an event-driven simulation package for modelling large networks of spiking neurons,” *Netw. Comput. Neural Syst.*, vol. 14, no. 4, pp. 613–627, 2003.

[82] C. D. Gilbert and W. Li, “Top-down influences on visual processing,” *Nat. Rev. Neurosci.*, vol. 14, no. 5, pp. 350–363, 2013.

[83] A. K. Engel, P. Fries, and W. Singer, “Dynamic predictions: oscillations and synchrony in top-down processing,” *Nat. Rev.*, vol. 2, no. 10, pp. 704–716, 2001.

[84] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[85] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, “Robust object recognition with cortex-like mechanisms,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 411–426, Aug. 2007.

[86] X. Hu, J. Zhang, J. Li, and B. Zhang, “Sparsity-regularized HMAX for visual recognition,” *PLoS One*, vol. 9, no. 1, p. e81813, 2014.

[87] C. Liu and F. Sun, “HMAX model: A survey,” *Proc. Int. Jt. Conf. Neural Networks*, vol. 2015–Septe, 2015.

[88] N. V. K. Medathati, H. Neumann, G. S. Masson, and P. Kornprobst, “Bio-inspired computer vision: Towards a synergistic approach of artificial and biological vision,” *Comput. Vis. Image Underst.*, vol. 150, pp. 1–30, 2015.

[89] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” *arXiv1502.03167 [cs]*, 2015.

[90] J. Kubilius, “Representation of configural information in the visual system,” KU Leuven, 2015.

[91] J. G. Snodgrass and M. Vanderwart, “A Standardized Set of 260 Pictures: Norms for Name Agreement, Image Agreement, Familiarity, and Visual Complexity,” *J. Exp. Psychol. Hum. Learn. Mem.*, vol. 6, pp. 174–215, Mar. 1980.

[92] A. Nguyen, J. Yosinski, and J. Clune, “Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images,” *arXiv1412.1897 [cs]*, 2014.

[93] Z. Zuo, B. Shuai, G. Wang, X. Liu, X. Wang, B. Wang, and Y. Chen, “Convolutional Recurrent Neural Networks: Learning Spatial Dependencies for Image Representation,” *IEEE Conf. Comput. Vis. Pattern Recognit. Work.*, pp. 18–26, 2015.

[94] Z. Zuo, B. Shuai, G. Wang, X. Liu, X. Wang, and B. Wang, “Learning Contextual Dependencies with Convolutional Hierarchical Recurrent Neural Networks,” *arXiv1509.03877 [cs]*, pp. 1–13, 2015.

[95] X. Pan, J. Shi, P. Luo, X. Wang, and X. Tang, “Spatial As Deep: Spatial CNN for

Traffic Scene Understanding," pp. 7276–7283, 2017.

[96] E. T. Rolls, "Invariant Visual Object and Face Recognition: Neural and Computational Bases, and a Model, VisNet," *Front. Comput. Neurosci.*, vol. 6, no. June, pp. 1–70, 2012.

[97] V. Goffaux and B. Rossion, "Faces are 'spatial'--holistic face perception is supported by low spatial frequencies.," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 32, no. 4, pp. 1023–1039, Jun. 2006.

[98] H. Halit, M. de Haan, P. G. Schyns, and M. H. Johnson, "Is high-spatial frequency information used in the early stages of face detection?," *Brain Res.*, vol. 1117, no. 1, pp. 154–161, Jun. 2006.

[99] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Am.*, vol. 74, no. 3, pp. 750–753, 1983.

[100] E. Zwicker, "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)," *J. Acoust. Soc. Am.*, vol. 33, no. 2, p. 248, 1961.

[101] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Am.*, vol. 8, no. 3, pp. 185–190, 1937.

[102] S. S. Stevens, "On the psychophysical law," *Psychol. Rev.*, vol. 64, no. 3, p. 153, 1957.

[103] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Process. Mag.*, no. November, pp. 82–97, 2012.

[104] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[105] A. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent

Neural Networks," *arXiv1303.5778 [cs]*, Mar. 2013.

[106] L. Caponetti, C. Buscicchio, and G. Castellano, "Biologically inspired emotion recognition from speech," *EURASIP J. Adv. Signal Process.*, vol. 2011, no. 1, p. 24, 2011.

[107] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and Accurate Recurrent Neural Network Acoustic Models for Speech Recognition," *arXiv1507.06947 [cs]*, 2015.

[108] D. Yu and L. Deng, *Automatic speech recognition: A Deep Learning Approach*, no. 1. Springer London Heidelberg New York Dordrecht ©, 2015.

[109] M. K. Al-Qaderi and A. B. Rad, "A Brain-inspired Multi-modal Perceptual System for Social Robots: An Experimental Realization," *IEEE Access*, vol. 6, pp. 35402–35424, 2018.

[110] M. Al-Qaderi and A. Rad, "A Multi-Modal Person Recognition System for Social Robots," *Appl. Sci.*, vol. 8, no. 3, p. 387, 2018.

[111] M. Al-Qaderi, E. Lahamer, and A. Rad, "A Two-stage Classifier Speaker Recognition System Based on Prosodic and Spectral Features via Fuzzy Inference Fusion," *Eurasip J. Audio, Speech, Music Process.*, no. (under review).

[112] R. Epstein, G. Roberts, and G. Beber, *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, 1st ed. Dordrecht: Dordrecht: Springer Netherlands, 2009.

[113] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, L. Wang, G. Wang, J. Cai, and T. Chen, "Recent Advances in Convolutional Neural Networks," *arXiv1512.07108 [cs]*, Dec. 2015.

[114] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015, pp. 36–45.

Appendix A.

A Brain-inspired Multi-modal Perceptual System for Social Robots: An Experimental Realization

Received June 6, 2018, accepted June 24, 2018, date of publication June 29, 2018, date of current version July 19, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2851841 (https://ieeexplore.ieee.org/ document/8400512)

MOHAMMAD K. AL-QADERI AND AHMAD B. RAD , (Senior Member, IEEE)
Autonomous and Intelligent Systems Laboratory, School of Mechatronic Systems Engineering, Simon Fraser University,
Surrey Campus, Surrey, BC V3T 0A3, Canada
Corresponding author: Ahmad B. Rad

Due to IEEE copyright restrictions, the full text of this article has been removed. The published version of “A Brain-inspired Multi-modal Perceptual System for Social Robots: An Experimental Realization” is available here:

Digital Object Identifier 10.1109/ACCESS.2018.2851841 (https://ieeexplore.ieee.org/ document/8400512)

Appendix B.

A Multi-modal Person Recognition System for Social Robots

applied sciences

MDPI

Article

A Multi-Modal Person Recognition System for Social Robots

Mohammad K. Al-Qaderi and Ahmad B. Rad *

Autonomous and Intelligent Systems Laboratory, School of Mechatronic Systems Engineering Simon Fraser University, Surrey, BC V3T 0A3, Canada; malqader@sfu.ca

* Correspondence: arad@sfu.ca; Tel.: +1-778-782-8512

Received: 16 January 2018; Accepted: 26 February 2018; Published: 6 March 2018

Abstract: The paper presents a solution to the problem of person recognition by social robots via a novel brain-inspired multi-modal perceptual system. The system employs spiking neural network to integrate face, body features, and voice data to recognize a person in various social human-robot interaction scenarios. We suggest that, by and large, most reported multi-biometric person recognition algorithms require active participation by the subject and as such are not appropriate for social human-robot interactions. However, the proposed algorithm relaxes this constraint. As there are no public datasets for multimodal systems, we designed a hybrid dataset by integration of the ubiquitous FERET, RGB-D, and TIDIGITS datasets for face recognition, person recognition, and speaker recognition, respectively. The combined dataset facilitates association of facial features, body shape, and speech signature for multimodal person recognition in social settings. This multimodal dataset is employed for testing the algorithm. We assess the performance of the algorithm and discuss its merits against related methods. Within the context of the social robotics, the results suggest the superiority of the proposed method over other reported person recognition algorithms.

Keywords: social robots; person recognition; multimodal machine perception; spiking neural network

1. Introduction

Recognizing people whom we have met before is, an indispensable attribute that is often taken for granted, yet playing a central role in our social interactions. It is suggested that humans can remember up to 10,000 faces (persons); though this is, an upper cognitive limit as, an average person remembers far less faces—around 1000 to 2000 different faces (persons) [1]. Humans are also remarkable in seamless and fast completion of various perceptual tasks, including object recognition, animal recognition, and scene understanding, to name a few. Acclaimed neurologist and author, the late Oliver Sacks, opined that the human brain is far less "prewired" than previously thought. In his highly readable and masterfully written book, "the Mind's Eye" [2], he talks about brain plasticity and how all the senses collectively contribute to form a perception of the world around us: "*Blind people often say that using a cane enables them to "see" their surroundings, as touch, action, and sound are immediately transformed into a "visual" picture. The cane acts as a sensory substitution or extension*". Within this setting, we mostly recognize people from their faces, though other characteristics such as voice, body features, height, and similar attributes often contribute to the recognition process.

In the context of the social robotics, it is very much desired that social robots, like humans, effortlessly distinguish familiar persons in their social circles without any intrusive biometric verification procedure. Consider how we recognize members of our family, co-workers, and close friends. Their faces, voices, their body shape, and features, etc., are holistically involved in the recognition process and the absence of one or more of these attributes usually do not influence the outcome of the recognition. There has been a large body of research that employs one or more biometric parameters for person re-identification for applications such as surveillance, security, or forensics systems. These features/parameters are derived from physiological and/or behavioral characteristics of humans, such as fingerprint, palm-print, iris, hand vein, body, face, gait, voice, signature, and keystrokes. Some of these features can be extracted non-invasively, such as face, gait, voice, odor, or body shape. In a parallel development, there are significant and impressive research studies that are focusing on face recognition which are generally non-invasive; however, it is important to distinguish the problem of person recognition from the face recognition.

Motivated by the above and noting that the problem of person recognition in social settings has not been investigated as widely as the related problems of person re-identification and face recognition; we propose a non-invasive multi-modal person recognition system that is inspired by the generic macrostructure of the human brain sensory cortex and is specifically designed for social human-robot interactions.

The rest of the paper is organized as follows: in Section 2, we outline the current state-of-art in multimodal person re-identification and face recognition systems. In Section 3, we present the detailed architecture and implementation of the proposed perceptual system for person recognition application. We will then include simulation studies and discuss the merits of the algorithm as opposed to other related methodologies in Section 4. We conclude the paper in Section 5.

2. Related Studies

The main thrust of the paper is to address the problem of person recognition in social settings. The problem presents new challenges that are absent in person re-identification scenarios such as surveillance, security, or forensics systems. Among these challenges are how to cope with changes in the general appearance of a subject due to attire change, extreme face and body poses, and/or variation in lighting. These challenges are further compounded by the fact that a concurrent presence of all biometric modalities is not always guaranteed. Moreover, the social robot is expected to complete the recognition task relatively fast (within the range of human reaction time in social settings). In addition, intrusive biometric verification procedure obviously is ruled out for social human-robot interaction scenarios. Nevertheless, multimodal biometric systems that non-invasively extract physiological and/or behavioral characteristics of humans, such as face, gait, voice, and body shape features have been reported to solve the person re-identification problem in social settings [3–6]. In such applications, the problem is treated as, an association task where a subject is recognized across camera views at different locations and times [7]. Due to the low resolution cameras and unstructured environments, these systems employ features such as, color, texture, and shape in order to identify individuals across a multi-camera network. However, these features are highly sensitive to variations in the subjects' appearance such as outfit or facial changes.

A person recognition system solely relying upon face recognition leads to erroneous detection if facial or environmental features change—such as growing beard, or substantial occlusion, variation in lighting, etc. There are also reported studies based on soft-biometric features that are non-invasive and are not much affected if the subject appears in different clothing [8,9]. Though, these methods rely upon a single biometric modality to extract specific auxiliary features. The performance of such systems is dramatically deteriorated in the absence of that dominant modality. A significant effort has been devoted to use face information as the main biometric modality in multimodal biometric recognition [9–11]. Within these classifications, multimodal biometric person recognition systems were proposed in [3,12]. These multimodal algorithms included a mixture of face, iris, fingerprint, and palm-print features. However, most of these studies also require other biometric features that cannot be extracted without the active cooperation of the subject, such as fingerprints, iris, and palm-prints. Hence, the overall multimodal biometric systems developed in most research studies fall within the invasive biometric system category. In contrast to the above methodologies, we introduce a non-invasive algorithm that does not require the cooperation of the subject as a requirement for its proper operation.

Since gait can be extracted non-intrusively from a distance, it is considered as, an important feature in developing person recognition systems. Gait is referred to as the particular manner in which a person walks and it is classified among the non-invasive attributes [13]. Zhou et al. [4] proposed a non-intrusive video-based person identification system based on integration of information from side face and gait features. The features are extracted non-invasively and fused at either feature level [4] or at the match score level [5]. In [3], the outputs of non-homogeneous classifiers, which are developed based on acoustic features from voice and visual features from face, are fused at the hybrid rank/measurement level to improve the identification rate of the system. Deep learning algorithms have also been used to address the problem of face recognition and action recognition, respectively [14,15]. Despite the fact that the above-mentioned studies are non-invasive multimodal biometric identification systems, the fusion methods that are employed in these systems require the concurrent presence of all biometric modalities for proper functioning, whereas the architecture that is reported in this paper relaxes this condition.

BioID [16] is a commercial multimodal biometric authentication system that utilizes synergetic computer algorithms to classify visual features (face and lip movements) and the vector quantifier to classify audio features (voice). The outputs of these classifiers are combined through different criteria to complete the recognition. In [8,17], facial information and a set of soft biometrics such as weight, clothes, and color were used to develop a non-intrusive person identification system, whereby the weight feature was estimated at a distance by the assessment of the anthropometric measurements that were derived from the subject's image captured by a standard resolution surveillance camera. The overall performance of the system was affected by the detection rate of the facial soft biometrics. In [13], the height, hair color, head, torso, and legs were used as complementary parameters along with the gait information for recognizing people. In order to improve the recognition rate of the system, the authors selected sets of these features along with gait information to be manually extracted from a set of surveillance videos. An intelligent agent-based decision-making person identification system was also reported in [18]. The system achieved a recognition rate of 97.67% when face, age, and gender information were used and a recognition rate of 96.76% when fingerprint, gender, and age modalities were provided to the system. A recent survey paper provides, an overview on using soft biometric (e.g., gender) as complementary information to primary biometrics (e.g., face) in order to enhance the performance of the person identification system [19]. Some researchers have applied multimodal biometrics systems to address related problems, such as action recognition [20], speaker identification [21], and face recognition [22].

The main shortcoming of these systems is that their different components require different time scales for proper operation, which limits their functionality in reaching decisions as compared to the human response time in different social contexts scenarios. For example, when the face is not detected due to extreme pose, partial occlusion, or/and poor illumination; the biometric features extracted from the face are not available and consequently the system fails to complete the recognition process. In contrast to these methods, the proposed approach overcomes this constraint by adjusting the threshold value of spiking neurons and exploiting available biometric features in order to compromise between the reliability of the decision and the natural perceptual time of the attended task (Section 3 of the paper).

Most of the aforementioned studies have been developed and discussed from a surveillance and security perspectives rather than the social human-robot interaction. Also, these person identification systems have been developed relying upon the combination of at least one dominant modality and a host of auxiliary biometrics or a mixture of invasive and non-invasive biometrics. Small number of research studies has tackled the problem of person recognition and face recognition in the context of cognitive developmental robotics [23,24]. We would like to emphasize that we present a person recognition algorithm incorporating multimodal biometrics features that is non-intrusive, is not affected by changes in appearance (i.e., outfit change), and works within the range of human social interaction rate (human response time). Moreover, all of the studies that were reported in multimodal biometric systems assume simultaneous presence of all the considered biometric features. This assumption is, however, relaxed in the proposed algorithm.

3. Architecture of the Person Recognition System

In this section, we present the architecture of the multi-modal person recognition algorithm in social settings. Figure 1 depicts the proposed system (Figure 1a) next to the architecture of the human/primate sensory cortex. Figure 1b shows a simplified architecture of the biological process as is widely accepted in neuroscience and psychophysics literature [25,26]. The architecture of the human sensory cortex is complex; it is thus naïve to claim, an exact reconstruction. Within this pretext, Figure 1a shows our interpretation, which is a much simpler functional "engineered replica" with a one-to-one correspondence to the biological system. In particular, although the pathways for each modality in the human sensory cortex are parallel; there are strong couplings between these pathways particularly after the primary receptive fields. In addition, the human sensory cortex is directly involved in motivation, memory, and emotions. In the proposed architecture, we have neither included the coupling effects of modalities nor have we considered emotions and memory. However, we have strictly adhered to the spirit of the multimodal parallel pathways. As depicted in Figure 1b,, an attended stimulus undergoes modality-specific processing (unimodal association cortex) before it converges at the higher level of the sensory cortex (multimodal association cortex) to form a perception [25–29].

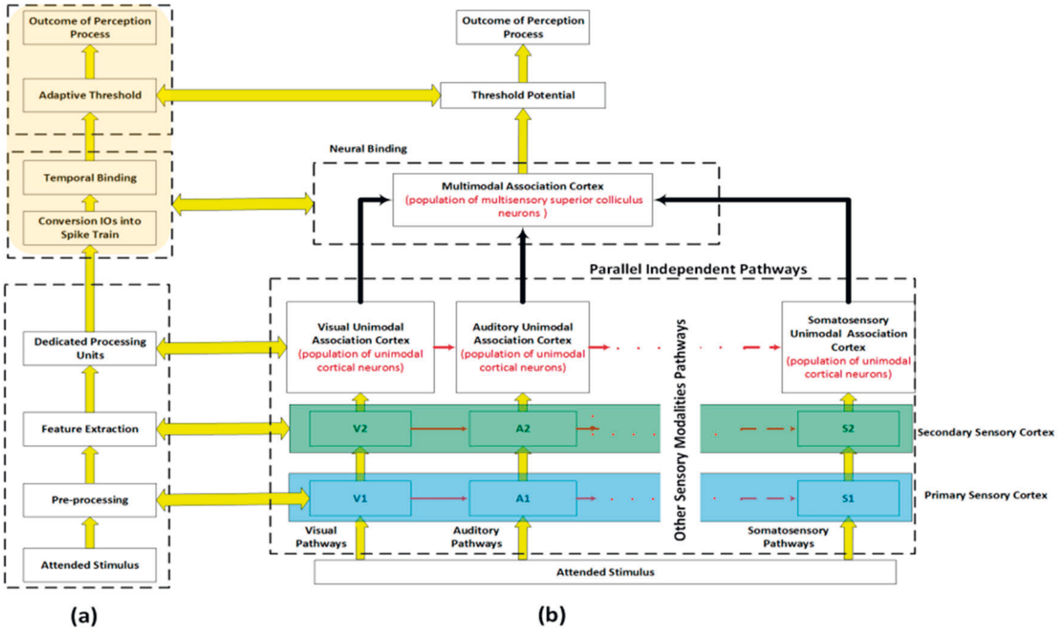
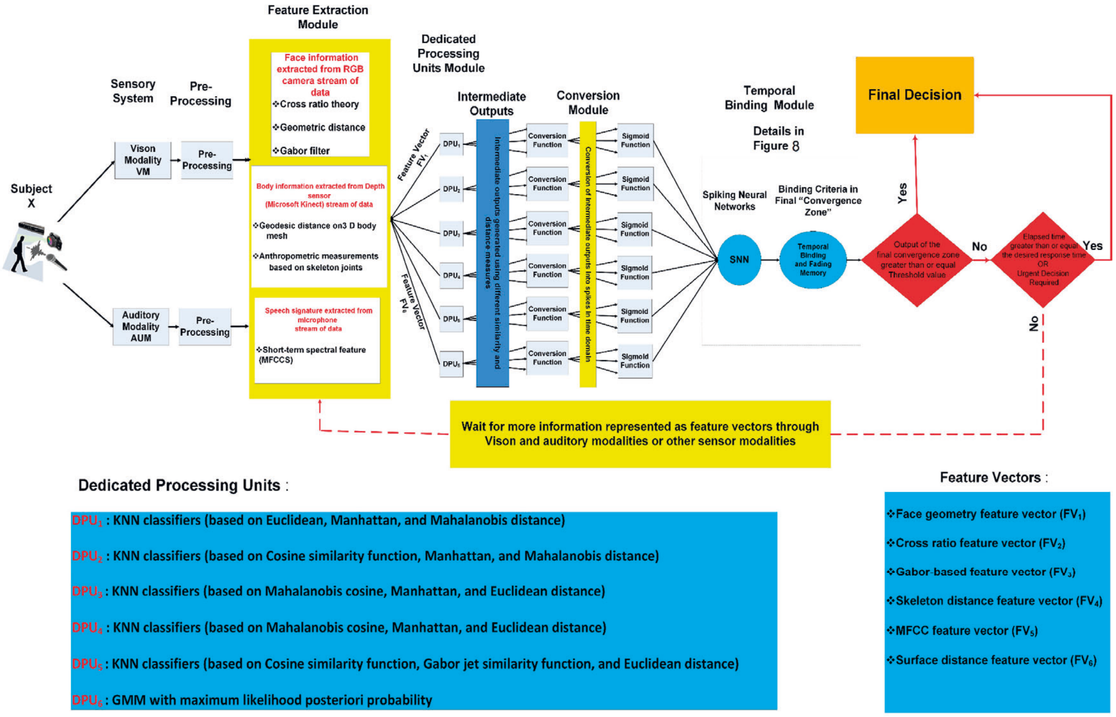

Outcome of Perception Process
Adaptive Threshold
Temporal Binding
Conversion IOs into Spike Train
Dedicated Processing Units
Feature Extraction
Pre-processing
Attended Stimulus
(a)
Outcome of Perception Process
Threshold Potential
Neural Binding
Multimodal Association Cortex
(population of multisensory superior colliculus neurons)
Parallel Independent Pathways
Visual Unimodal Association Cortex
(population of unimodal cortical neurons)
Auditory Unimodal Association Cortex
(population of unimodal cortical neurons)
Other Sensory Modalities Pathways
Somatosensory Unimodal Association Cortex
(population of unimodal cortical neurons)
V2
A2
S2
Secondary Sensory Cortex
V1
A1
S1
Primary Sensory Cortex
Visual Pathways
Auditory Pathways
Somatosensory Pathways
Attended Stimulus
(b)

Figure 1. The proposed framework of brain-inspired multimodal perceptual system for social robots.

An attended stimulus to each of the visual, auditory, and somatosensory (touch, pressure, pain, etc.) systems undergoes a preprocessing and a feature extraction module, i.e., V1 and V2 in visual, A1 and A2 in auditory, and S1 and S2 in somatosensory pathways. When excited by a real-world stimulus, the corresponding neural systems of the human sensory system (vision, auditory, and tactile) map the stimulus's attributes to available modalities. A similar structure is employed in the proposed system as elaborated in Figure 2 whereby each sensor modality and even different types of information within a sensor modality are processed in parallel and through independent processing pathways at the early stages of the perception process (feature extraction modules and dedicated processing units). The outputs of these independent pathways (intermediate outputs) converge at the higher level (temporal binding) to form the final outcome of the perception process. Also, in the biological system, an attended stimulus is mapped by population of neurons distributed across and within the cortical hierarchy, the binding or perceptual grouping is accomplished by synchronization of neural firings among population of neurons that form the cell assembly [26]. Then, the integration of the outputs of these cell assemblies in parallel with the search for the best match of the attended pattern, within the library of representations stored in memory, and perceive the attended stimulus. The findings from neuroscience and psychophysics suggest that the formation of cell assemblies is controlled by the following principles: (1) population of neurons in a specific cell assembly must have similar receptive field properties, (2) each cell assembly maps one feature or quality of the attended stimulus, and (3) population of neurons in the same cell assembly fire in temporal synchrony with each other.

We have incorporated these principles in the proposed architecture, as shown in Figure 1a and further elaborated in Figure 2. The first principle is depicted by connecting each sensory modality to a dedicated pre-processing and feature extraction module (corresponding to primary and secondary sensory cortex), which in turn generates a set of feature vectors representing different attributes of the attended stimulus. Feeding each of these feature vectors to its corresponding Dedicated Processing Unit (DPU) satisfies the second principle. Each feature vector is then processed by its respective DPU, which in turn, contributes to the production of the Intermediate Outputs (IOs). In the rest of this paper, we refer to the outputs generated by each DPU as simply IOs. The variation in the processing time that is required to generate the IOs and the availability of biometric modalities in the sensory system streams are handled by the binding modules. Psychophysics and psychological research studies suggest that the face recognition process uses two type of information: configural information and featural information, which are available at low and high spatial frequency, respectively. The former is used in early stage of recognition process and requires less processing time whereas the latter is used to refine and rectify the recognition process at the later stage and requires more processing time [30,31]. IOs will be transformed into temporal spikes in order to be processed by the temporal binding system. At the last stage, the output of the temporal binding module is compared with, an adaptive threshold setting to either complete the perception process or to wait for more information from other sensor modalities. This adaptive threshold is controlled by two factors: the desired reliability of the final outcome and how fast a decision is required. In some scenarios, a fast response is more important than, an accurate response; thus, the threshold will be reduced to accommodate such scenarios. For example, in the context of social robots, the natural (in the human sense) and relatively fast response is more desirable than, an accurate but slow response [32]. In some situations, when, an urgent decision is required, humans process a real-world stimulus by exploiting the most discriminant feature [33]. In such cases, a fast processing route is selected as the outcome at final convergence zone even the threshold value is not satisfied. However, in other situations when accurate response is more important than fast response, humans may take longer time and look for other cues to perceive reliably and accurately. The proposed framework accommodates both conditions by incorporating, an adaptive threshold. The proposed architecture is customized to address the person recognition problem in social contexts, as shown in Figure 2. However, the same architecture may be adapted to solve other perceptual tasks that are vital in social robotics, including but not limited to, object recognition, scene understanding, or affective computing.


Subject
X
Sensory System
Pre-Processing
Vison Modality VM
Pre-Processing
Auditory Modality AUM
Pre-Processing
Feature Extraction Module
Face information extracted from RGB camera stream of data
❖Cross ratio theory
❖Geometric distance
❖Gabor filter
Body information extracted from Depth sensor (Microsoft Kinect) stream of data
❖Geodesic distance on3 D body mesh
❖Anthropometric measurements based on skeleton joints
Speech signature extracted from microphone stream of data
❖Short-term spectral feature (MFCCS)
Dedicated Processing Units Module
Feature Vector FV_1
Feature Vector FV_6
DPU_1
DPU_2
DPU_3
DPU_4
DPU_5
DPU_6
Intermediate Outputs
Intermediate outputs generated using different similarity and distance measures
Conversion Module
Conversion Function
Conversion of intermediate outputs into spikes in time domain
Sigmoid Function
Temporal Binding Module
Details in Figure 8
Spiking Neural Networks
SNN
Binding Criteria in Final "Convergence Zone"
Temporal Binding and Fading Memory
Output of the final convergence zone greater than or equal Threshold value
Yes
No
Final Decision
Elapsed time greater than or equal the desired response time OR Urgent Decision Required
Yes
No
Wait for more information represented as feature vectors through Vison and auditory modalities or other sensor modalities

Dedicated Processing Units :

DPU_1 : KNN classifiers (based on Euclidean, Manhattan, and Mahalanobis distance)

DPU_2 : KNN classifiers (based on Cosine similarity function, Manhattan, and Mahalanobis distance)

DPU_3 : KNN classifiers (based on Mahalanobis cosine, Manhattan, and Euclidean distance)

DPU_4 : KNN classifiers (based on Mahalanobis cosine, Manhattan, and Euclidean distance)

DPU_5 : KNN classifiers (based on Cosine similarity function, Gabor jet similarity function, and Euclidean distance)

DPU_6 : GMM with maximum likelihood posteriori probability

Feature Vectors :

❖Face geometry feature vector (FV_1)

❖Cross ratio feature vector (FV_2)

❖Gabor-based feature vector (FV_3)

❖Skeleton distance feature vector (FV_4)

❖MFCC feature vector (FV_5)

❖Surface distance feature vector (FV_6)

Figure 2. Sophistication of the proposed framework for person recognition task in social settings. DPU: Dedicated Processing Unit; KNN: K-Nearest Neighbors; GMM: Gaussian Mixture Models.

For the specific problem of person recognition, the architecture employs auditory as well as vision modalities. Though, the framework readily allows for the integration of additional modalities (tactile, olfaction) for other applications. As shown in Figure 2, when the sensory system (vision, auditory) is excited by a real-world stimulus, the corresponding receptive field system generate a map for the available stimulus's attributes in a parallel manner. We refer to this stage in the architecture as pre-processing and feature extraction modules. It is well documented in psychophysics and neuroscience research where not only the processing of different sensor modalities is performed by independent routes of processing, but also different kinds of information within the same modality are processed by independent processing paths [26,34,35].

The capability of the perceptual system to finalize the perceptual task (person recognition in this study) in the absence of concurrent availability of all sensor modalities is utilized by using the spiking neurons in the binding modules (Section 3.4). For example, if the subject's face is not available, then the binding module may use other available cues, such as body features, speech features, or both of them in order to finalize the recognition process within a reasonable response time (within the norm of the human response/reaction time). As depicted in Figure 2, a compromise between the reliability of the outcome and the requirement of quick response (in the order of human natural reaction) is achieved by, an adaptive threshold (more on that in Section 3.4). We describe each module in more details in the rest of this section.

3.1. Front-End Sensors and Preprocessing

In order to address the person recognition problem, the visual and auditory pathways are employed. An RGB camera and a three-dimensional (*3D*) depth sensor (i.e., Kinect sensor) may be applied to capture the image and the corresponding depth information (vision modality/pathway), and a microphone could be employed to process the voice of a subject (auditory modality/pathway). The Kinect sensor as a *3D* multi-stream sensor captures a stream of colored pixels; depth information associated with these colored pixels, and positioned sound. The data streams from the RGB camera, *3D* depth sensor, and the microphone are processed via standard signal and image preprocessing (filtering and noise removal, thresholding, segmentation, etc.) to be prepared for the feature extraction module (Figure 2). In this study, however, such preprocessing is not required as we extract the input data from three databases that already provide preprocessed data.

3.2. Feature Extraction

In this section, we introduce the feature extraction stage, which is analogous to the primary and secondary sensory cortex in the human brain. The input of this module is the preprocessed data stream from the vision and auditory modules and its outputs are distinct feature vectors that will be processed by the respective classifiers as computational models for the DPUs (Figure 2).

The target application of the proposed person recognition system is social robotics. One of the most important and desirable attributes of the social robots is the ability to recognize individuals in various settings and scenarios, including challenging scenarios whereby one or more sensor modalities are temporary not available such as in vision system whereby lighting is inadvertently changed, or subjects change their outfits. Many reported methodologies have difficulties in coping with such unstructured settings.

In order to configure the perceptual system to person recognition tasks, three types of features, which are available in the data streams of auditory and vision system, need to be extracted. These features are categorized in three groups: the first group is based on spatial relationship, referred to as configural features; the second group is the appearance-based feature, which relies upon texture information; and the third group of feature is a voice-based feature which relies upon short-term spectral feature.

3.2.1. Vision-Based Feature Vectors

The vision-based feature vectors consist of two groups of feature vectors: The configural features group and the appearance-based feature group. Most of the feature vectors in the configural features are available early in the recognition process due to their relatively less computational requirements. On the other hand, the extraction of the appearance-based feature group is computationally expensive and is available later in the recognition process. This is also compatible with psychology and neuroscience findings that spatial information is processed early in the perception process and provides a coarse categorization scheme for, an attended stimulus.

The Configural Features Group

The group consists of four feature vectors. The first feature vector is represented by the ratios of the Euclidian distances among the geometric position of a set of fiducial points on a face. These fiducial facial points are detected by “OpenFace”;, an open source software for facial landmark detector [36]. The second feature vector is based on a cross ratio of the projection lines that are initiated from the corners of the polygon constructed from a set of predefined fiducial points on a face image. The third feature vector is constructed by computing the Euclidian distance among a set of selected skeleton joint positions. The fourth feature vector in this group is the surface-based feature, which is generated by computing the geodesic distances between the projections of selected pairs of skeleton joints on the point cloud that represent, an individual’s body. It is worth mentioning that these feature vectors are purposely selected as they are easy to calculate and available early in perception process. The main purpose for these feature vectors are to limit the search scope and provide shortlisted candidates for the attended subject by biasing the top-ranked spiking neurons (see Section 3.4 for more details).

The first feature vector in the configural group consists of eight facial feature ratios, as shown in the Appendix A (Table A1). Despite the simplicity of this geometric descriptor, it can be shown that they generate comparable performance in face clustering with respect to other feature vectors that describe face appearance such as EigenFace and Histogram of Oriented Gradients [37].

The second feature vector was constructed by employing the cross ratio theorem, which is a widely applied object and shape recognition algorithm in computer vision [38]. The cross ratio value stays invariant under geometric projection operations such as translation, rotation, and scaling changes [39]. The cross ratio of four collinear points A, B, C, and D in a line L, as shown in Figure 3, is given by:

$$CR_L(A, B, C, D) = \frac{|\overline{AC}| \cdot |\overline{BD}|}{|\overline{BC}| \cdot |\overline{AD}|}$$

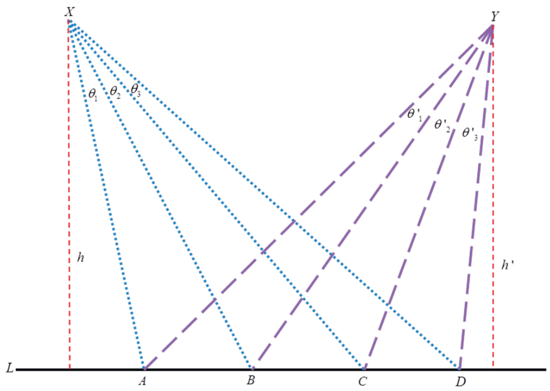


Figure 3. The cross ratio relationship of two viewpoints.

The same cross ratio, CR_L, can also be expressed as ratio of the projection lines XA, XB, XC, and XD. By using the fact that the XAB triangle area can be calculated using the formulas: $\frac{1}{2} * h * AB = \frac{1}{2} * XA * XB * sin\theta_1$ and some algebraic manipulation, the cross ratio from point X, can be expressed as a function of the line segments as in (1) or as a function of projection angles as in (2), where h is the distance between the focus and the line AB, as depicted in Figure 3.

$$CR_X(A,B,C,D) = \frac{|\overline{AC}|\cdot|\overline{BD}|}{|\overline{BC}|\cdot|\overline{AD}|} \tag{1}$$

$$CR_X(\theta_1,\theta_2,\theta_2) = \frac{sin\ (\theta_1+\theta_2)\cdot sin\ (\theta_2+\theta_3)}{sin\theta_2\cdot sin\ (\theta_1+\theta_2+\theta_2)} \tag{2}$$

Since the cross ratio value is independent of changes in the viewpoint, the cross ratio of the same four collinear points A, B, C, and D in a line L from point Y can be expressed in the same way as point X as in (3) and (4).

$$CR_Y(A,B,C,D) = \frac{|\overline{AC}|\cdot|\overline{BD}|}{|\overline{BC}|\cdot|\overline{AD}|} \tag{3}$$

$$CR_Y(\theta\prime_1,\theta\prime_2,\theta\prime_3) = \frac{sin\ (\theta\prime_1+\theta\prime_2)\cdot sin\ (\theta\prime_2+\theta\prime_3)}{sin\theta\prime_2\cdot sin\ (\theta\prime_1+\theta\prime_2+\theta\prime_3)} \tag{4}$$

hence, $CR_X(\theta_1,\theta_2,\theta_3) = CR_Y(\theta\prime_1,\theta\prime_2,\theta\prime_3)$. The reader may refer to [39] for detailed proof. Where X and Y are two different viewpoints, $\{\theta_1,\theta_2,\theta_3\}$, $\{\theta\prime_1,\theta\prime_2,\theta\prime_3\}$ represent the projection angles from point X and Y respectively as shown in Figure 3.

The same principle is applied to measure the similarity of polygons that are constructed by selecting five points from the pre-defined fiducial points on a face image, as shown in Figure 4b. One fiducial point is used as the basis point and the other four must be non-collinear fiducial points to represent the polygon. The cross ratio of this polygon is regarded as the basis of similarity measure that is not affected by translation, scaling, rotation, and illumination. More details about the cross ratio for face recognition can be found in [39]. The set of five cross ratios is calculated by switching the basis point to one of the polygon's corners, and the cross ratio values are obtained using (1) to (4).

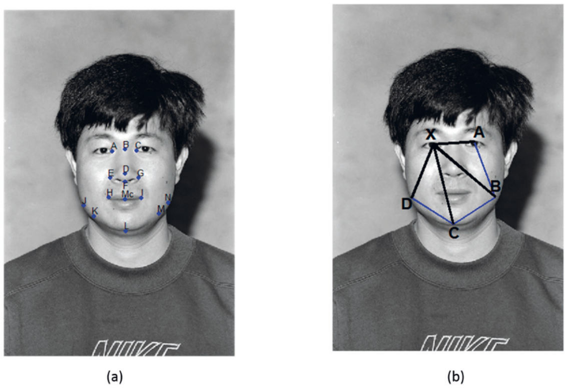


Figure 4. (**a**) Selected fiducial points on image from the FERET database which are used to construct the configural feature vector, (**b**) The cross ratio projection based on a polygon constructed from five fiducial points on face image from the FERET database.

The third feature vector in this group is the skeleton-based feature. The combination of the distances between the selected skeleton joints, shown in Figure 5a, are used to generate this feature vector, as described in Table A2 and depicted in Figure 5b (The reader may refer to Appendix B for further details).

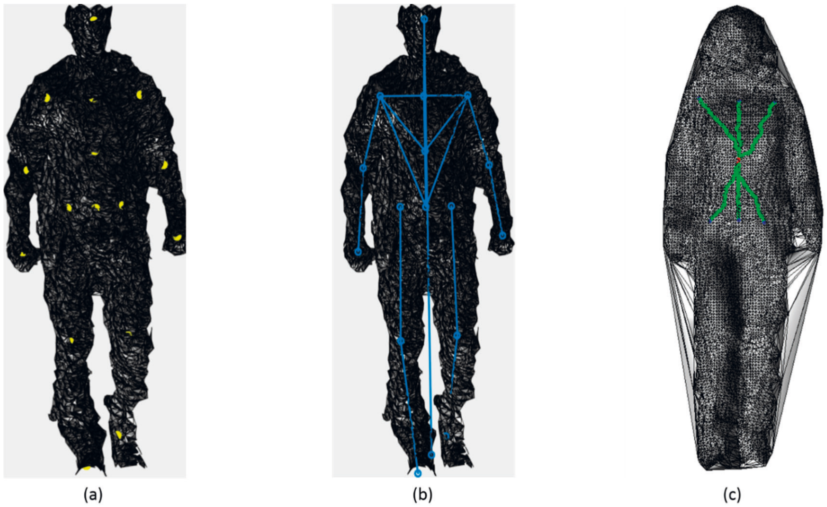


Figure 5. Selected skeleton joints, geodesic and Euclidean distance among them; (**a**) Projection of skeleton joints on the three-dimensional body point cloud, (**b**) Euclidian distance of selected skeleton segments, and (**c**) Sample of geodesic paths used in constructing surface-based feature vector.

The surface-based feature vector is the last vector in the configural features group. This feature vector is computed using the combination of geodesic distances among the projection of selected skeleton joints on the three-dimensional body point cloud. First, the selected pairs of the skeleton joints, which do not usually lie on the point cloud, are projected on the associated closest point on the three-dimensional body mesh, which is generated from the point cloud. The pair of the projection points is used to initiate the fast-marching algorithm that provides a good approximation of the shortest geodesic path between two points on the surface. The fast-marching algorithm uses a gradient descent of the distance function to extract a good approximation of the shortest path (geodesic), as given by the Dijkstra algorithm [40]. Figure 5c depicts, an example of geodesic distances used in constructing the surface-based feature vector. The selected geodesic distances that used to construct the surface-based feature vector are described in Table A3 (Appendix B).

The Appearance-Based Feature Group

The appearance-based feature consists of a set of multi-scale and multi-orientation Gabor filter coefficients extracted from the face image at fiducial points. The authors are aware of the availability of stronger descriptors like Scale-Invariant Feature Transform (SIFT) [41] and Speeded-Up Robust Features (SURF) [42], both of which can be used to generate feature vectors with high discrimination power. However, our intention is to process configural information early in the computation through, an independent processing path in order to limit the number of candidates of, an attended stimulus that can be refined further by the information available in the appearance-based feature. This interpretation is also compatible with the findings in neuroscience [34] and psychology [30] on human object and face recognition, suggesting that spatial information is used in early stages of the recognition.

The regional facial appearance patterns are normally extracted by the Gabor filter as a set of multi-scale and multi-orientation coefficients that represent the appearance-based feature vector. The Gabor filter may be applied to the whole face or to specific points on the face [43,44]. Extraction of Gabor filter coefficients is computationally expensive due to convolution integral operation; therefore, in order to speed up the computation, the Gabor filter coefficients are only computed at the fiducial points shown in Figure 4a. The two-dimensional (*2D*) Gabor filter centered at (0, 0) in the spatial domain can be expressed as in (5):

$$G(x, y, \xi_x, \xi_y, \sigma_x, \sigma_y, \theta) = \frac{1}{\sqrt{\pi\ \sigma_x\sigma_y}} e^{-\frac{1}{2}} \left[(\frac{R_1}{\sigma_x})^2 + (\frac{R_2}{\sigma_y})^2 \right] e^{j(\xi_x\ x + \xi_y\ y)} \tag{5}$$

where $R_1 = x\ cos\theta + y\ sin\theta$ and $R_2 = -x\ sin\theta + y\ cos\theta$, ξ_x and ξ_y are spatial frequencies, σ_x and σ_y are the standard deviation of, an elliptical Gaussian along the x and y axes, and θ represents the orientation. The Gabor filters have a plausible biological model to resemble the primary visual cortex. Physiological studies suggest that cells in the primary visual cortex usually have, an elliptical Gaussian envelope with, an aspect ratio of 1.5–2.0; thus, one can infer the following relation [45]:

$$\xi_x = \omega\ cos\ \theta, \xi_y = \omega\ sin\ \theta$$

Daugman [46] suggests that simple and complex cells in the primary visual cortex have plane waves propagating direction along the short axis of the elliptical Gaussian envelope. By defining the aspect ratio $r = \sigma_y/\sigma_x$ and assuming that the minimum value of aspect ratio is 1, the Gabor filter has, an elliptical Gaussian envelope and the plane wave's propagating direction along the $x - axis$, which is the shortest in case of $r > 1$, can be expressed as (6):

$$G(x, y, \omega, \sigma, r, \theta) = \frac{1}{\sqrt{\pi\ r\ \sigma}} e^{-\frac{1}{2}} \left[\left(\frac{R_1}{\sigma}\right)^2 + \left(\frac{R_2}{r\sigma}\right)^2 \right] e^{j(\omega\ R_1)} \tag{6}$$

where $\sigma = \sigma_y$ and $r = \sigma_y/\sigma_x$. Given, an input image I, the response image of the Gabor filter can be computed using the convolution operation defined as in (7). We convolve the image I with every Gabor filter kernel in the Gabor filter banks centered at the pixels specified by the fiducial points.

$$z = \sum_x \sum_y I(x, y) G(x' - x, y' - y, \omega, \sigma, r, \theta) \tag{7}$$

where $G(x' - x, y' - y, \omega, \sigma, r, \theta)$ is Gabor filter kernel centered at (x', y'). $I(x, y)$ is the intensity value of the image I at (x, y) location. The performance of the Gabor filter response in face recognition and classification tasks is highly affected by the parameters that are used in construction of the Gabor Kernel bank [44]. One of the well-known Gabor filter banks that is widely used in many computer vision applications especially object and face recognition tasks is the "classical bank". The "classical bank" is characterized by eight orientations and five frequencies with $f_{max=0.25}\ pixel^{-1}$, $f_{ratio} = \sqrt{2}$, $\sigma = \sigma_x = \sigma_y = \sqrt{2}$, *and* $\phi = 0\ radians$. Many previous studies have been devoted to addressing the problem of finding the Gabor filter parameters, which have optimum performance on the recognition tasks [43,47–49]. In this study, we adopted the Gabor filter parameters suggested by [44]. The author of that paper claims that the following parameterization of Gabor filter extracts the most discriminant information for recognition tasks. The suggested parameters are: eight orientations, six frequencies (instead of 5) with narrower Gaussian width ($\sigma_x = \sigma_y = 1$ instead of $\sqrt{2}$ that is used in classical setting). The rest of the parameters were set the same as in the "classical bank" setting. The Gabor filter bank responses given in (7) consist of real and imaginary parts that can be represented as magnitudes and phases components. Since the magnitudes vary slowly with the position of fiducial points on the face, where the phases are very sensitive to them, we used only the magnitudes of the Gabor filter responses to generate the appearance-based feature vector. Hence, we have 48 Gabor coefficients for each fiducial point on the face. The selected set of Gabor filter kernels and responses are depicted in Figure 6; for demonstration, we selected one scale {1}, two orientations $\{\frac{\pi}{8}, \frac{5\pi}{8}\}$, and three frequencies $\{\frac{0.25}{(\sqrt{2})^5}, \frac{0.25}{(\sqrt{2})^3}, \frac{0.25}{\sqrt{2}}\}$ to create Figure 6a,b. Figure 6a shows the magnitude of Gabor filtered kernels that were used to compute these coefficients at the fiducial points. Figure 6b depicts the magnitude of Gabor filter responses on a sample image from the FERET database (FERET database will be further discussed in Section 4).

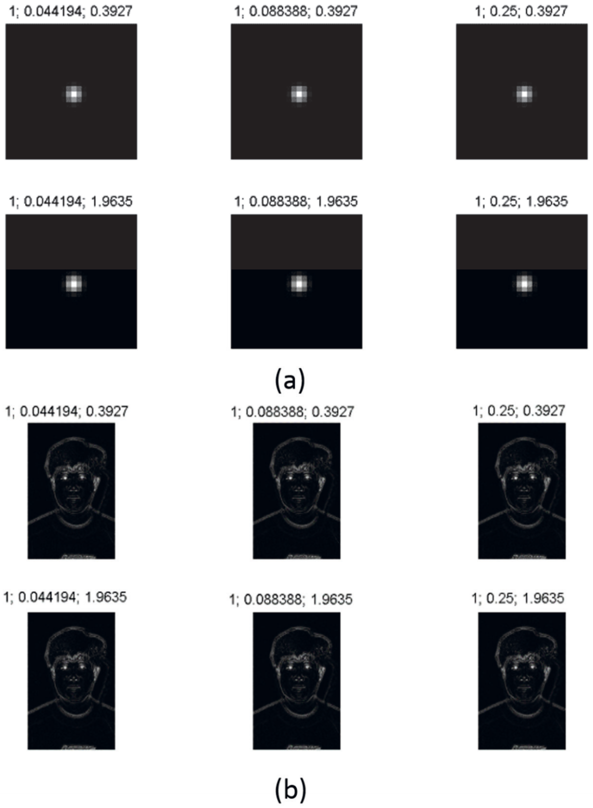


Figure 6. (**a**) Magnitude of Gabor filtered kernels at one scale {1}, two orientations $\{\frac{\pi}{8}, \frac{5\pi}{8}\}$, and three frequencies $\{\frac{0.25}{(\sqrt{2})^5}, \frac{0.25}{(\sqrt{2})^3}, \frac{0.25}{\sqrt{2}}\}$. (**b**) magnitude of Gabor filter responses on a sample image from the FERET database at one scale {1}, two orientations $\{\frac{\pi}{8}, \frac{5\pi}{8}\}$, and three frequencies $\{\frac{0.25}{(\sqrt{2})^5}, \frac{0.25}{(\sqrt{2})^3}, \frac{0.25}{\sqrt{2}}\}$.

3.2.2. Voice-Based Feature Vector

The voice-based feature vector is computed based on the short-term spectral, specifically, the so-called mel-frequency cepstral coefficients (MFCCs). We opted for MFCCs for many reasons: (1) MFCCs are easy to extract compared to other speech features, such as voice source features, prosodic feature, and spectro-temporal features; (2) MFCCs require relatively less amount of speech data to be extracted; and (3) MFCCs is text and language independent. Thus, MFCCs feature vector fits the nature of the person recognition for the social HRI where a real-time response and text-independent speech signature are crucial for user acceptance of social robot. A modular representation of MFCCs feature vector extraction is shown in Figure 7.

MFCCs feature vector is computed based on a widely accepted suggestion that the spoken words cover a frequency range up to 1000 Hz. Thus, MFCCs use linearly spaced filter at low frequency below 1000 Hz and logarithmic spaced filter at high frequency above 1000 Hz. In other words, the filter-bank is condensed at the most informative part of the speech frequency (more filters with narrow bandwidths below 1000 Hz) and lengthy-spaced filter-bank is applied at higher frequencies. As depicted in Figure 7, the first step in the extraction process is to pre-emphasize the input speech signal by applying filter as in (8).

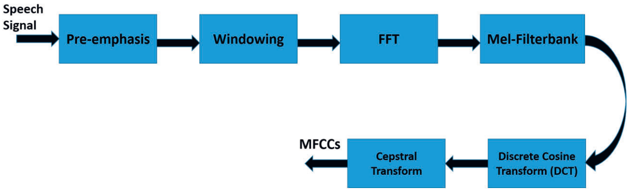


Figure 7. Modular representation of mel-frequency cepstral coefficients (MFCCs) feature extraction process.

$$Y(n) = X(n) - a \ * X(n-1) \tag{8}$$

where $Y(n)$ is pre-emphasized speech signal, $X(n)$ is the input speech signal, and a pre-emphsized factor can be any value in the interval $[0.95, \ 0.98]$. In the next step (Windowing), the pre-emphasized speech signal $Y(n)$ is multiplied by smooth window function, here, we used Hamming windows, as in (9).

$$W(n) = 0.54 - 0.46 * cos(\frac{2\pi n}{N-1}), \quad 0 \leq n < N-1 \tag{9}$$

The resultant time-domain signal is converted to frequency domain by applying the well-known Fast Fourier Transform (FFT). The frequency range in the resultant FFT spectrum is very wide and fluctuated. Thus, the filter-bank that is designed according to Mel scale is applied in order to get the global shape of the FFT spectrum magnitude which is known to contain the most distinctive information for speaker recognition. The MFCCs are obtained by applying logarithmic compression and discrete cosine transform, as in (10). The discrete cosine transform converts log Mel spectrum into the time domain.

$$C_n = \sum_{m}^{M} [\log S(m)] cos \left[\frac{\pi n}{M} (m - \frac{1}{2}) \right] \tag{10}$$

where $S(m)$, $m = 1, 2 \ldots, M$ is output of, an M-channel filter-bank, n is the index of the cepstral coefficient. In this study, we retained the 12 lowest C_n excluding 0th coefficient.

3.3. Dedicated Processing Units and Generation of the Intermediate Outputs

As explained in the previous section, given a sequence of facial images, *3D* mesh, and speech data for a person in various social settings, six feature vectors are extracted and considered to participate in the perception of, an attended stimulus in order to recognize the person from different subjects in the database. The rationale for the choice of the six features is that the algorithm is architecturally and is functionally inspired by the human perceptual system. It is established that humans have limited channel capacity of processing the information from their sensory system. This capacity varies in the range of five to nine according to a seminal research study [50]. These feature vectors are: face geometry feature, cross ratio feature, skeleton feature, surface distance feature, appearance-based feature, and speech-based feature (Section 3.2). These features vectors are fed to DPUs in order to generate IOs. Selection of possible computational models of these DPUs is problem dependent and relies upon the perceptual task that needs to be addressed, as discussed in previous section.

3.3.1. Dedicated Processing Units for Vision-Based Feature Vectors

For the vision-based feature vector, we adopt classifiers that use various similarity and distance measures to represent their respective DPUs. These classifiers generate scores that evaluate how similar or close a subject is from those in the gallery. The interpretation of the best match relies on the types of distance and similarity measures that were used to generate these scores. For instance, in the case of various distance measures, such as L2Norm, L1Norm, Mahalanobis distance, and Mahalanobis Cosine; the minimum score represents the best match (please refer to Appendix A for more details). Whereas in cases where IOs are calculated using similarity measures, such as Cosine Similarity; the maximum score represents the best match. However, in order to unify these measures, such that the maximum score represents the best match; distance measures, they are further modified as (11).

$$IO_{jk} = \frac{\log\left(D^*_{j1}+1\right)}{\log\left(D^*_{jk}+1\right)} \tag{11}$$

where $D^*_{j1} \leq D^*_{j2} \leq \ldots\ldots.. \leq D^*_{jk}$, represent various distance measures, IO_{jk} is a unified score value representing how much the jth subject from the test set match or close to the kth subject from gallery set. It can be seen from (11) that the smallest distance yields a score value (IO) or a confidence value closer to one, while the largest distance value produces a very small score (IO) or a confidence value that is close to zero. These unified scores are then converted into spike times compatible with the inputs of neurons in the spiking neural network (SNN) at the next stage of hierarchical structure. For each subject in the test set, each feature vector participating in encoding the attended stimulus is processed by its respective DPU. In this study, DPUs are selected to be K-Nearest Neighbors (K-NN) classifiers which use a combination of three of the following similarity and distance measures: L2Norm, L1Norm, Mahalanobis distance, Mahalanobis Cosine, and Cosine Similarity as detailed in the architecture shown in Figure 2. Each DPU generates three matrices by adopting three of the aforementioned similarity and distance measures to compute scores for its associated feature vectors in the evaluation set against the corresponding feature vectors in the gallery set. However, only for the face appearance feature vector, the Gabor Jet Similarity measure of each subject in the evaluation set, is computed against the corresponding face appearance feature vector in the gallery set using (12) and (13).

$$Sim^i_a(J, J\prime) = \frac{\sum_{k=1}^{N} a_{ki}a'_{ki}}{\sqrt{\sum_{k=1}^{N} a^2_{ki} \sum_{k=1}^{N} a'^2_{ki}}} \tag{12}$$

$$Sim_{face} = \sum_{i=1}^{L} Sim^i_a(J, J\prime) \tag{13}$$

where $Sim^i_a(J, J')$ is the similarity between two jets, J and J' associated with ith fiducial points on the face of the subject, a_{ki} is the amplitude of kth Gabor coefficient at ith fiducial points. N is the number of wavelet kernels. Sim_{face} represents the total similarity between the two faces as the sum of the similarities over all the fiducial points as expressed in (13).

3.3.2. Dedicated Processing Units for Voice-Based Feature Vector

For the speech-based feature vector, MFCCs (mel-frequency cepstral coefficients) are used as, an input to K-NN classifier with the aforementioned distance measures. Also, we used MFCCs that extracted from speech data of all of the speakers in the training data (gallery set) to create speaker-independent world model or a well-known universal background model (UBM). The UBM is estimated by training M-component GMM with the popular expectation–maximization (EM) algorithm [51]. The UBM represents speaker-independent distribution of the feature vectors. Here, we use 32-compnenent GMM to build the UBM. The UBM is represented by a GMM with 32-compnents, as denoted by λ_{UBM}, that characterized by its probability density function as (14).

$$p(\overrightarrow{x}|\lambda) = \sum_{i=1}^{M} w_i p_i(\overrightarrow{x}) \tag{14}$$

The model is estimated by the weighted linear combination of D-variate Gaussian density function $p_i(\overrightarrow{x})$, each parameterized by a mean $D \times 1$ vector, μ_i, mixing weights, which is constrained by $w_i \geq 0$, $\sum\limits_{i=1}^{M} w_i = 1$, and a $D \times D$ covariance matrix, Σ_i as (15).

$$p_i(\overrightarrow{x}) = \frac{1}{2\pi^{D/2}|\Sigma_i|^{1/2}} exp\{\frac{1}{2}(x-\mu_i)'(\Sigma_i{}^{-1})(x-\mu_i)\} \tag{15}$$

The purpose of training the UBM is to estimate the parameters of 32-component GMM, $\lambda_{UBM} = \{w_i, \mu_i, \Sigma_i\}_{i=1}^{M}$, from the training samples. The next step is to estimate specific GMM from UBM-GMM for each speaker in the gallery set using maximum a posteriori (MAP) estimation. The key difference between estimating the parameters of UBM and estimating the specific GMM parameters for each speaker is that the UBM uses standard iterative expectation-maximization (EM) algorithm for parameter estimation. On the other hand, specific GMM parameters are estimated by adapting the well-trained parameters in the UBM to fit a specific speaker model. Since the UBM represents speaker-independent distribution of the feature vectors, the adaptation approach facilitates the fast scoring, as there is a strong coupling between speaker's model and the UBM. It should be noted that all or some of the GMM's parameters ($\lambda_{UBM} = \{w, \mu, \Sigma\}$ can be adapted by MAP. Here, we adapted only the mean μ to represent specific speaker's model. Now, Let us assume a group of speakers $s = 1, 2, 3, \ldots,\ S$ represented by GMMs $\lambda_s = \lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_S$. The goal is to find the speaker identity $\hat{s}$ whose model has the maximum a posteriori probability for a given observation $X_k = \{x_1, \ldots, x_T\}$ (MFCCs feature vector). We calculate the posteriori probability of all of the observations $X_k = X_1, X_2, X_3, \ldots, X_K$ in probe set against all of the speakers models $\lambda_s = \lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_S$ in gallery set as (16). As s and k vary from 1 to number of speakers in the gallery set and the number of utterances in probe set, respectively, the result from (16) is $S \times K$ matrix, namely $IO\_FV_{voice\_based}$. This matrix represents the IOs that are generated from speech-based feature vector and it will be integrated with other matrices that represent IOs generated from vision-based feature vectors.

$$IO\_FV_{voice\_based}|_{\{s,k\}} = P_r(\lambda_s|X_k) = \frac{p(X_k|\lambda_s)}{p(X_k)} P_r(\lambda_s)\Bigg|_{\substack{1 \leq s \leq S \\ 1 \leq k \leq K}} \tag{16}$$

Assuming equal prior probabilities of all the speakers, the terms $P_r(\lambda_s)$ and $p(X_k)$ are constant for all speakerx, thus both terms can be ignored in (16). Since each subject in the probe set is represented as $X_k = \{x_1, \ldots, x_T\}$, thus by using logarithmic and assume independence between observations, calculation of $IO\_FV_{voice\_based}|_{\{s,k\}}$ can be simplified as (17).

$$IO\_FV_{voice\_based}|_{\{s,k\}} = \sum_{t=1}^{T} \log p(x_k^t|\lambda_s)|_{\substack{1 \leq s \leq S \\ 1 \leq k \leq K}} \tag{17}$$

Each feature vector generates IOs matrices, which provide a degree of support for each class in the gallery set based on several measures within the same feature vector. Also, IOs matrices that generated from different feature vectors provide a degree of support for each class in the gallery set in a complementary manner. The weight contribution of the IOs generated from the same feature vector to the final output is less than that of IOs generated from different feature vectors when they are integrated in the Spiking Neural Networks (SNN). This will be further discussed in the next section.

The next problem is to distinguish a subject x from the M subjects in the gallery set. Several IOs matrices are calculated for vision-based feature vector to be integrated with IOs matrices generated from the speech-based feature vector. Each matrix takes the size of a $M \times C$ matrix and its name is formatted based on the feature vector that generated it. The matrix name is read as

$IO\_FV_{name\ of\ feature\ vector}$. For example, the matrices that are describing the resultant IOs based on the skeleton feature vector should read as $IO\_FV_{skeleton}$, where M represents the number of subjects in the gallery set and C is the number of samples in test set.

$$\begin{aligned} IO\_FV_{name\ of\ feature\ vector} &= \begin{bmatrix} io_{11} & \cdots & io_{1C} \\ \vdots & \ddots & \vdots \\ io_{M1} & \cdots & io_{MC} \end{bmatrix}, where\ IOV_j = [io_{j1},\ io_{j2},\ \ldots,\ io_{jc}] \\ &= \begin{bmatrix} IOV_1 & \cdots & IOV_M \end{bmatrix}^T \end{aligned}$$

where IOV_j is the IOs vector, io_{jk} represents how much the jth subject from the test set match or close to the kth subject from gallery set. This score is associated with a specific feature vector and is generated based on a certain distance measure that is specified by the name of the matrix.

3.4. Temporal Binding via Spiking Neural Networks

It is known that humans interact with their environment by processing the available information through multisensory modality streams over time with fading memory property. The same process is emulated here. In the context of this algorithm, fading memory implies that the effect of stimuli excitation (represented by IOs) deteriorates moderately if it is not reinforced or refreshed. We implement this feature through the Leaky Integrate-and-Fire neuron (LIF) model [52] to manifest the integration of IOs that are generated in the previous stage in the hierarchical architecture of the proposed system. Figure 8 depicts one block of the spiking neural network (SNN) that is used to perform the integration process. The overall SNN that is used to integrate the information from various biometric modalities is constructed by laterally connecting *N* blocks from the circuit, as shown in Figure 8, where *N* represents number of subjects in gallery sets.

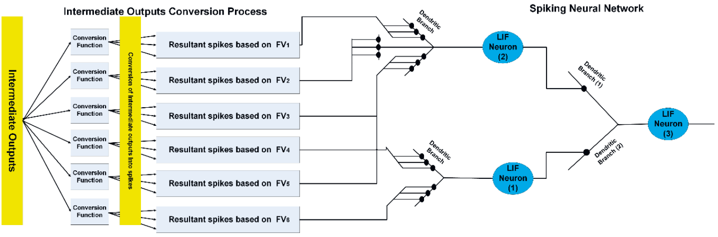


Figure 8. Spiking neural network (SNN) circuit and it dendritic structure that are used as a block to construct the overall SNN. LIF: leaky integrate-and-fire.

The IOs vectors are fed to LIF neurons in SNN by means of pre-synaptic input spikes, as shown in Figure 8. IOs vectors, which are generated based on different feature vectors, are fed to independent branch in the dendritic tree. On the other hand, IOs vectors that are generated based on same feature vectors are fed to same branch in the dendritic tree. Inspired by neuroscience research studies [53,54], we suggest that the effect of presynaptic inputs on postsynaptic potential is either sublinear, super linear, or linear. The effects sum sub-linearly, linearly, or super-linearly if they are delivered to the same dendritic branch (within-branch) and sum linearly if they are delivered to different dendritic branches (between-branch). Equation (18) describes the dynamic of postsynaptic potential of LIF neuron. The dynamic of this neuron can be described as follows: initially at time *t* = 0, *Vm* is set to *Vinit*. If *Vm* exceeds the threshold voltage *Vthresh*, then it fires a spike and it is reset to *Vreset* and held there for the length *Trefact* of the absolute refractory period. The total response of postsynaptic potential due to different presynaptic inputs within-branch (Syn_{WB}) and between-branch (Syn_{BB}) is computed using (18) to (20).

$$\tau_m \frac{dV_m}{dt} = -(V_m - V_{resting}) + R_m \cdot (I_{syn}(t) + I_{noise}) \quad (18)$$

where $\tau_m = C_m \cdot R_m$ is the membrane time constant, R_m is the membrane resistance, $I_{syn}(t)$ is the current supplied by the synapses, I_{noise} is a Gaussian random variable with zero mean and a given variance noise, *Vm* is the membrane potential of LIF neuron, *Vinit* is the initial condition for *Vm* at time $t = 0$, and *Vthresh* is the threshold value. If *Vm* exceeds *Vthresh,* then a spike is emitted, *Vreset* is the voltage to reset *Vm* to after a spike, and $V_{resting}$ is the membrane potential of LIF neuron at no activity.

$$Syn_{WB} = \sum_{i=1}^{K} \alpha_i \; Sigmoid(IO_i) \tag{19}$$

$$Syn_{BB} = \sum_{i=1}^{K} \alpha_i \; IO_i \tag{20}$$

where IO_i represents the total input to the *i*th dendritic branch, α_i is the *ith* dendritic branch weight, *K* is the number of dendritic branches. Note that the sigmoid function is one possible choice of synaptic integration function within-branch and can be replaced with other functions, such as hyperbolic tangent sigmoid function.

It can be noted from (19) and (20) that the balanced IOs that are delivered to same dendritic branch will sum as follows: (1) small IOs will sum nearly linearly, (2) around average IOs will sum super-linearly, (3) large IOs will sum sub-linearly. Unbalanced IOs fed to the same branch generate near-linear summation over the entire range of IOs intensities. Moreover, IOs that are delivered to independent branches will sum linearly for all of the combinations of IOs intensities. Synaptic integration in dendritic tree of pyramidal neuron was experimentally proved to demonstrate similar behavior to the aforementioned forms of summations [55]. These forms of summations provide a tradeoff between error variance and error bias. Sub-linear summation of within-branch IOs, in case of large IOs, reduces the error variance by not exaggerating the effect of one aspect of the measure at the expense of other measures in deriving the final outcome. In addition, the linearly weighted aggregation of between-branch IOs reduces error bias by means of exploiting various attributes in deriving the final outcome.

As shown in Figure 8, the integration of IOs is performed using SNN in time domain to emphasize the temporal binding with fading memory criteria. The IOs represent various scores of confidence; each one of them provides a degree of support for each subject in the gallery set according to a certain aspect of measure and based on a specific biometric modality. These scores are introduced to SNN as presynaptic inputs by means of spikes fired at different times. As described in the previous section, all of the IOs are unified such that high score is equivalent to best match. In order to introduce the IOs to LIF neurons, the IOs are converted to spike times using (11) such that a high IO is equivalent to early firing time. Hence, the neuron which fires first represents the best candidate of the attended subject (from the gallery set). As LIF neurons receive early spikes, which correspond to high degree of support, their membrane potential U increases instantaneously. Once the membrane potential U of one of these neurons crosses the threshold value V_{thresh}, the neuron fires a spike and all neurons participating in the process are reset to V_{reset}. The neuron which fires a spike first, which we refer to as the winner neuron, represents the best candidate of the attended subjects, and the attended subject is labeled with class number assigned to that neuron.

The threshold value of the neurons in the SNN controls both the reliability of the perception outcome and the allowed for the perception time of the attended task. A LIF neuron with a high threshold value implies that it will not fire until high intensity presynaptic inputs are delivered to its dendrite branches. These presynaptic inputs may be not available due to the absence of some biometric features or the need for more processing time. Thus, a compromise between the reliability and the reasonable perception time can be achieved by controlling the threshold value, according to a specific scenario of social interaction. As IOs are introduced to LIF neurons in parallel (Figure 8) via presynaptic inputs, one very high IO may drive a neuron to fire a spike and finalize the perception process. This sheds light on the superior feature of this model, such that one biometric feature with high discriminant power may be enough to finalize the perception process. This feature replicates the ability of humans to recognize odd features very quickly [56]. In the face perception and recognition, humans focus on distinctive features, which correspond to very high IOs in this algorithm so that other features may not need to be used.

Another vital property of this model is the alleviation of the computational cost in the perception process. As one of the neurons in the final layer fires a spike, all of the neurons that are participating in the perception of attended stimulus are reset and held at that state for a certain time. Early spikes correspond to IOs that carry high discriminant power and consequently provide high degree of support for particular neuron to be the winner neuron and represent the best candidate of attended subject; however, the neuron receiving the earliest spike is not necessarily the winner neuron. In some cases, a neuron receives a spike later, but is reinforced immediately with other spikes that will drive its potential to threshold value and consequently fire a spike before other neurons, which were received the earliest spikes but were not immediately reinforced with other spikes. As shown in Figure 9a, even though neuron 1 receives a spike prior to neuron 2, neuron 2 fires a spike earlier than neuron 1. It can be seen from Figure 9a that the membrane potential of neuron 1 had started increasing earlier than the membrane potential of neuron *2*, but because neuron 2 received a spike and reinforced immediately with another spike, its membrane potential increased dramatically and had fired before the membrane potential of neuron 1 reached the threshold value. Figure 9b shows the case that one IO, which corresponds to a very early input spike, is large enough to drive the neuron's potential to threshold value and fires a spike. One can tentatively conclude that a neuron fires a spike either by a very high IO, corresponding to very early spike that is sufficiently large to drive a neuron's potential to threshold, or by more than one high or moderate IO, representing a monotonically decreasing function and corresponding to spikes that are reinforced each other in time domain. The number of neurons which represent the final layer of SNN (i.e., outputs of SNN) equals the number of subjects in the gallery set. Thus, the first neuron fired among these neurons represents the best candidate of attended subject and the attended stimulus is labeled with the number of that neuron.

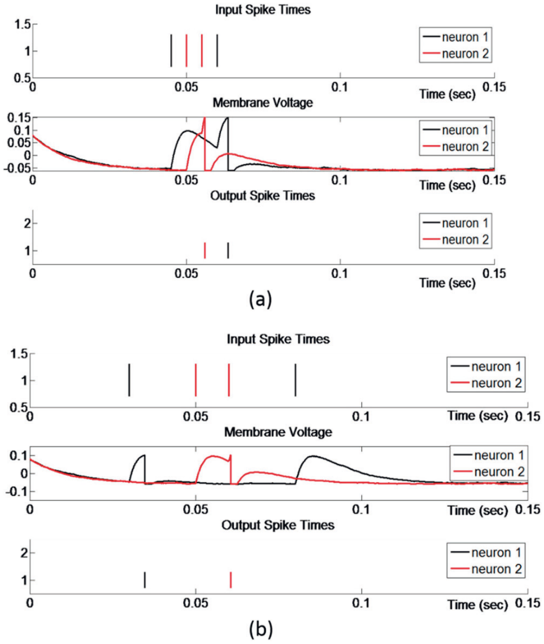
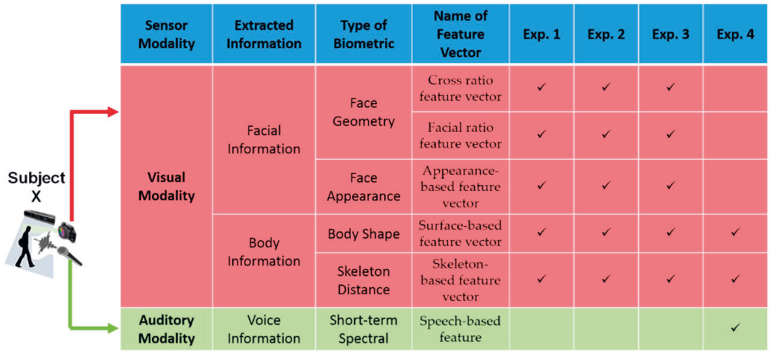


Figure 9. (**a**) The earliest spike is not sufficiently large to drive the neuron's potential to threshold and evokes a spike, (**b**) The earliest spike is sufficiently large to drive the neuron's potential to threshold and evokes a spike.

4. Experimental Results

In this section, we present the experimental results to evaluate the performance of the person recognition algorithm in social settings. We have included four sets of simulation studies for person recognition to demonstrate the performance of the person recognition algorithm. The biometrics that have been extracted from visual and auditory modalities are presented in three groups, as shown in Figure 10. The biometrics that have been selected to identify a subject in each of the four scenarios are illustrated in Figure 10.



Sensor Modality	Extracted Information	Type of Biometric	Name of Feature Vector	Exp. 1	Exp. 2	Exp. 3	Exp. 4
Visual Modality	Facial Information	Face Geometry	Cross ratio feature vector	✓	✓	✓	
			Facial ratio feature vector	✓	✓	✓	
		Face Appearance	Appearance-based feature vector	✓	✓	✓	
	Body Information	Body Shape	Surface-based feature vector	✓	✓	✓	✓
		Skeleton Distance	Skeleton-based feature vector	✓	✓	✓	✓
Auditory Modality	Voice Information	Short-term Spectral	Speech-based feature				✓

Figure 10. The Biometrics that have been discussed in the four experiments.

4.1. Generation of Multi-Modal Data Set

Our first challenge was that the available public datasets are generally unimodal, and as such, do not fit to the requirements of the multimodal perception. We resolved this problem by creating a new dataset from merging of the three datasets: FERET [57], TIDIGITS [58], and RGB-D [59]. FERET database contains a total of 14,126 facial images of 1199 individuals and 364 duplicate sets of facial images. TIDIGITS is a speech dataset that was originally collected at Texas Instruments Inc. (Dallas, TX, USA) The TIDIGITS corpus contain 326 speakers (111 men, 114 women, 50 boys and 51 girls), with each pronouncing 77 digit sequences. The RGB-D is a new database that was created by Barbosa et al. for the purpose of person re-identification studies based on information from *3D* depth sensor. In this dataset, depth information has been obtained for 79 individuals with four scenarios: frontal view of person walking normally (Walking 1 group), frontal view of person walking slowly and avoiding obstacles (Walking 2 group), walking with stretched arms (Collaborative group), and back view of person walking normally (Backward group). Five synchronized information for each person namely, RGB images, foreground mask, skeleton, *3D* mesh, and the estimated floor were collected in, an indoor environment, whereby the individuals were at least two meters away from the *3D* depth sensor.

In order to provide the individual in RGB-D database with facial images from a diverse group across ethnicity, gender, and age, we randomly selected 79 subjects from FERET database. Then, we used only frontal view images, which included frontal images at different facial expressions (*fb* image), different illuminations (*fc* image). Also, some subjects in the database wore glasses on and/or pull their hair back. The duplicate set contains frontal images of a person which was taken on a different day over one year, and for some individuals more than two years had elapsed between their first frontal images and the duplicate ones. The number of frontal facial images for each subject in the selected set varies from two to eight images. These 79 subjects were randomly assigned to subject in RGB-D database when considering that female subjects from FERET database are assigned to female subjects from RGB-D.

In order to complement the new dataset with speech data; we selected 23 subjects from women group in TIDIGITS dataset and assigned them randomly to female subjects in the new dataset, the rest of subjects in the new dataset were assigned with speech data from men group in TIDIGITS dataset.

The new dataset provides facial information, speech utterances, and the aforementioned information that is available on RGB-D database. The facial information is extracted from FERET database which provides facial frontal images with some differences such as changes in facial expression, change in illumination level, and variable amount of time between photography sessions. Also, RGB-D database provides skeleton and depth information that not affected by changing the outfits of the subjects and their bodies poses. On the other hand, TIDIGITS provide speaker signature when the subject is not in the field of view of robot's vision system. It is important to note that the state-of-art face detection and recognition algorithms fail to provide quick detection and have low recognition rate when the face is angled or far from the camera, or when the face is partially occluded, and/or the illumination is poor. However, these situations are common in social HRI scenarios. In such cases, other biometrics features, such as body information and speech signature, can be used to compensate missing facial information and recognize, an individual. These characteristics of the new database fit the requirements of the human-robot interactions in social settings where robust long-term interaction is a crucial factor for the success of the system.

The new (integrated) dataset has been partitioned into two sets, namely, training (gallery) and evaluation (probe) sets, as described in experiments 1 to 4. The gallery set was used to build the training model and the evaluation set was used for testing. The evaluation set is comprised of unseen data, not used in the development of the system. It is important to emphasize that the chronological order of the data capture was considered in constructing the evaluation set. Thus, some of the images in the evaluation set was chosen to be duplicate I and duplicate II, implying that they were taken at different dates, spanning from one day to two years. By using duplicate I and II images in constructing the evaluation sets, we ensured that the evaluation set represented closely scenarios that are appropriate for long-term HRI in social settings. The performance of the proposed architecture was evaluated in four experiments. Since, the data set has 79 subjects, thus the overall SNN was constructed from 79 circuits, as shown in Figure 8. In this SNN, all of the LIF neurons number 3 are connected laterally and all blocks have the same dendritic structure shown in Figure 8.

4.2. Experiment 1

For each subject in the probe set, two facial images, *fb* image and its duplicate I image, were selected from the FERET database. In addition, two out of five frames from each of skeleton information and *3D* mesh body information were selected randomly from Walking 1 group in the RGB-D database. The rest of the samples in the FERET and RGB-D databases were used to construct the training set. Some subjects in the FERET database had only two facial images. In this case, one was used for training and the other for evaluation. Five feature vectors were constructed, as described in Section 3. Three of the feature vectors represent facial information, including the facial geometry feature vector, cross ratio feature vector, and appearance-based feature vector. The rest of the feature vectors, namely the skeleton feature vector and the surface-based feature vector, represent the body information of the attended subject. IOs generated based on these features were converted into spike times and normalized to range from *zero to* 150 ms, prior to being fed to LIF neurons in SNN, as shown in Figure 8. The SNN was constructed and simulated using the neural Circuit (CSIM) simulator [60]. The parameters of LIF neurons were set as follows: the weight synapses of neuron 1 and neuron 2 were equal and set at 2000×10^{-9}. The weight synapses of neuron 3 were set as follows: the weight synapse of dendritic branch one was set to 2500×10^{-9} and weight synapse of dendritic branch two was set to 2000×10^{-9}, $V_{thresh} = 0.15$, $V_{reset} = -0.067$, $V_{reseting} = 0$, $C_m = 5 \times 10^{-8}$, $V_{init} = 0.08$, $R_m = 1 \times 10^6$, $T_{refact} = 0.0025$, $I_{noise} = 50 \times 10^{-9}$, $I_{sys}(\mathrm{t})$ represents the input current supplied by the synapses, i.e., the outputs from the conversion process of IOs into input spike times. These input spike times were set in the range from *zero to* 150 ms. This selection is compatible with the natural human perception of time. The SNN were simulated for 150 ms. As described in Section 2, the first neuron that fires a spike represents the best candidate of the attended subject *x* from the gallery set. The overall SNN was constructed from 79 circuit blocks, as shown in Figure 8. Therefore, the total number of LIF neurons was 237. The recognition rates were calculated at two stages in the hierarchical structure of the SNN, namely stage 1 and stage 2. Stage 1 consists of the list of neurons, labeled as neuron 1 and neuron 2; stage 2 was represented by the list of neurons labeled as neuron 3. The recognition rate that was calculated from the list of neurons labeled as neuron 1 was based on body information; the recognition rates that were calculated from the list of neurons labeled as neuron 2 expressed a recognition rate based on facial information or voice information. Neuron 2 may use face geometry, face appearance, voice-based feature, or all of them in order to fire a spike. The same applies to neuron 1, which may use geodesic distances, skeleton distances, or both, in order to drive its potential to the threshold and consequently evoke a spike. The overall recognition rates were calculated based on neuron 3, which may use facial information, body information, voice information, or a combination of them. Cumulative match curves (CMCs) show the probability that the correct match of classification is found in the *N*, the most likely candidates, where *N* (the rank) is plotted on the *x*-axis. CMCs provide the performance measure for biometric recognition systems and have been shown to be equivalent to the ROC of the system [61]. The recognition result was averaged over ten runs; the cumulative match curves (CMCs) were plotted for these recognition results and are shown in Figure 11a–c.

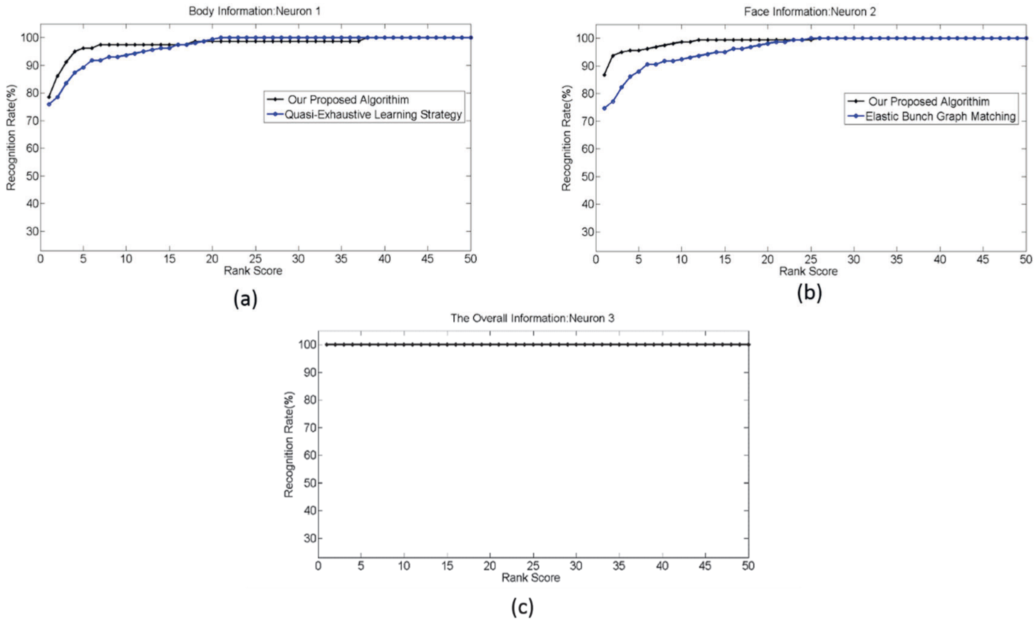


Figure 11. (**a**) Cumulative match curves (CMC) based on body information calculated on Walking 1 group from the RGB-D database, (**b**) CMC based on face information calculated on fb and duplicate I images from the FERET database, and (**c**) CMC based on temporal binding of face and body information where body information is evaluated as described in Figure 11a and face information is evaluated as described in Figure 11b.

4.3. Experiment 2

In this experiment, the probe set was constructed as follows: for body information, we used the collaborative group from the RGB-D database as the training set and two frames out of five from Walking 2 group as the probe set. For facial information, the probe set was constructed from *fb* image and duplicate *II* image. The rest of the samples in the FERET database was used to construct the training set. It can be noted that the probe and training sets were constructed in this manner to demonstrate the performance of the system in a real-world scenario where the enrolment process of the attended subject happened when the subject's posture was different from that of the recognition process. All the other configurations of SNN were similar to the experiment 1. The recognition result was averaged over ten runs. The cumulative match curves (CMCs) were plotted for these recognition results and are shown in Figure 12a–c. The overall recognition rate is degraded as result of using different groups from the RGB-D database for training and evaluation. Hence, the same person is represented in one posture in gallery set and a different posture in the probe set. Another reason for the performance degradation is the use of the duplicate *II* image set to construct the probe set for the face information. This is a huge challenge for the state-of-the-art face recognition algorithms due to changes in illumination, aging, and facial expressions. Nevertheless, the proposed algorithm works reasonably well.

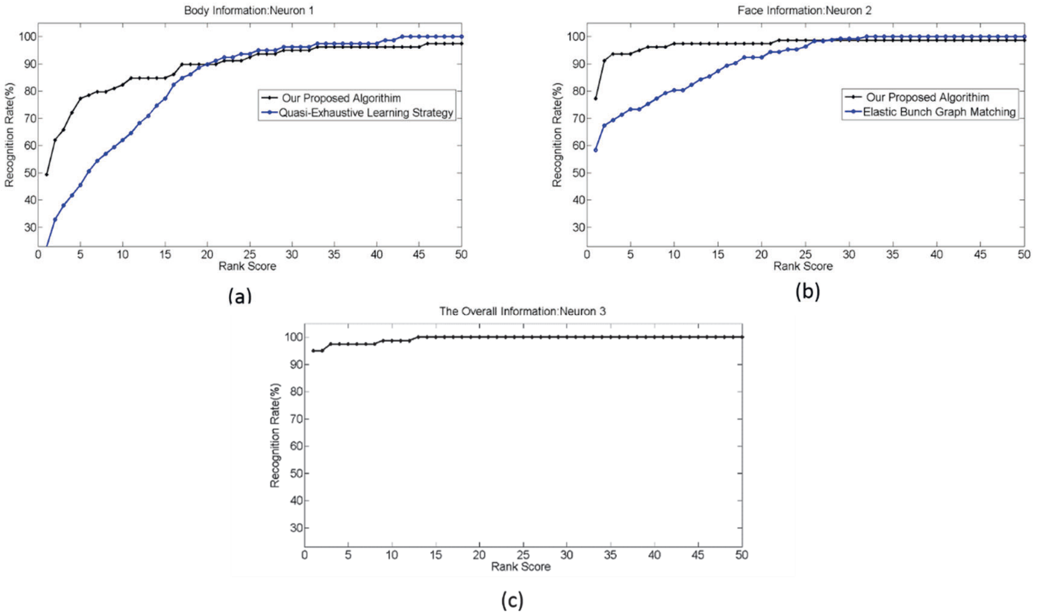


Figure 12. (**a**) CMC based on body information calculated on Walking 2 group vs collaborative group from the RGB-D database, (**b**) CMC based on face information calculated on fb and duplicate II images from the FERET database, and (**c**) CMC based on temporal binding of face and body information where body information is evaluated as described in Figure 12a and face information is evaluated as described in Figure 12b.

4.4. Experiment 3

We emulated a real-world scenario of HRI in social settings where biometric modalities that represent person identity are not concurrently available due to the sensor limitation or the occlusion of some parts of the person. To replicate this scenario, we converted the IOs generated from the body information into temporal spikes in range of 0–150 ms while the IOs that are generated from the face information were converted into temporal spikes in the range of 30–150 ms In this way, we made the body information available before the face information. This scenario replicates a situation where a person can be identified from his skeleton and body shape before face biometric modalities are available. Here, we assumed that the back view of the attended person is captured by the RGB-D sensor at the beginning of the recognition process and after a short time the attended person turned toward the camera in such a way that the face information becomes available. Hence, two frames out of five from the backward group in the RGB-D database are used to construct the probe set. For facial information, the probe set was constructed from *fb* image and duplicate II image, the same as in experiment 2. The rest of the samples in the FERET and RGB-D databases were used to construct the training set. All of the configurations of SNN are similar to the first experiment. The recognition result was averaged over ten runs, and the cumulative match curves (CMCs) were plotted for these recognition results, as shown in Figure 13a–c. The recognition rates are still good, despite the fact that biometric modalities are available at different times. We have not seen any other algorithm that copes with this scenario.

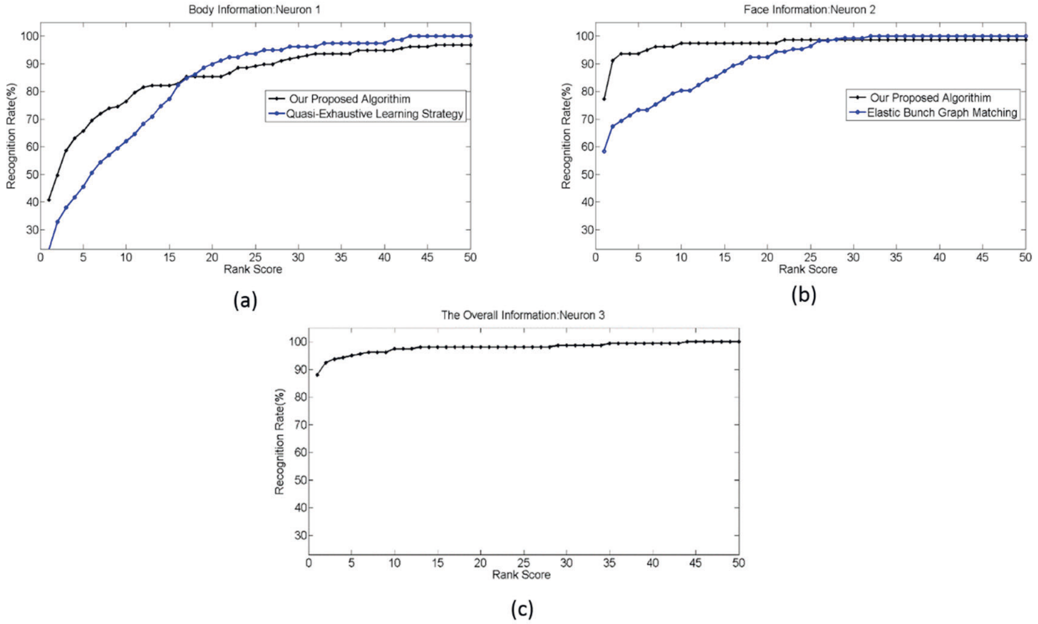


Figure 13. (**a**) CMC based on body information calculated on Backward group from the RGB-D database, (**b**) CMC based on face information calculated on fb and duplicate II images from the FERET database, (**c**) CMC based on temporal binding of face and body information where body information is evaluated as described in Figure 13a and face information is evaluated as described in Figure 13b.

4.5. Experiment 4

In this experiment, we emulated another challenging scenario of HRI in social settings when a subject's face is not detected either due to distance between the robot and the subject or due to titled viewing angle of the camera and the head orientation. However, we assume that some utterances from subject's speech can be captured by robot's auditory system, as well as *3D* mesh for subject's body is available in robot's vision stream of data. In this scenario, the subject's speech signature and his/her body information are available. Here, we assumed that the audio signal is recorded first and the voice activity detector is applied such that only the voice signal is fed to speech feature extraction module. Also, we assumed that speech utterances of the attended person are captured by a microphone at the beginning of the recognition process, and after a short time, the attended person shows in camera's view facing opposite way such that back view of body information becomes available. Thus, for each subject, two frames out of five from the backward group in the RGB-D database were used to construct the probe set for body information. The rest of the samples in the RGB-D database was used to construct the training set. For speech signature, the probe set was constructed by selecting seven utterances (each utterance in range of 1 to 1.7 s duration) out of 77 utterances from TIDIGITS database for each subject. The rest of the samples in the TIDIGITS database was used to construct the training set. Despite the fact that short speech utterances (such as the ones used in constructing the probe set for speech signature) reduce the recognition rate, we used them in our implementation to demonstrate its reasonable performance in this challenging HRI scenario. All of the configurations of SNN are similar to the first experiment. The recognition result was averaged over ten runs, and the cumulative match curves (CMCs) were plotted for these recognition results, as shown in Figure 14a–c. The recognition rates are still good, despite the fact that biometric modalities are available at different times and only two of them are available. We have not seen any other algorithm that copes with this scenario.

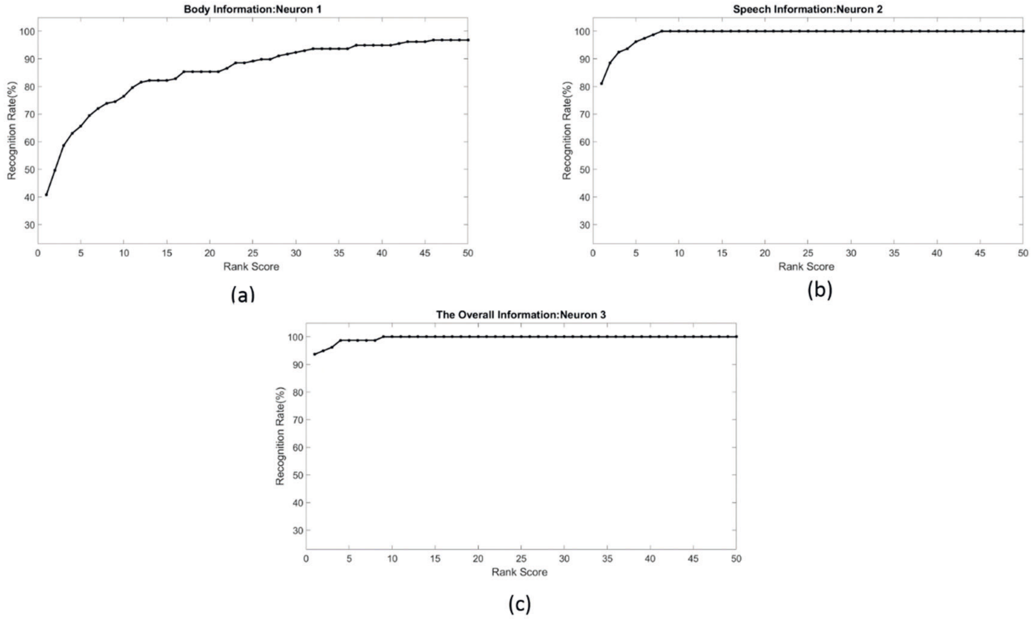


Figure 14. (**a**) CMC based on body information calculated on Backward group from the RGB-D database, (**b**) CMC based on speech information calculated on selected utterances from the TIDIGTS database, and (**c**) CMC based on temporal binding of speech and body information where body information is evaluated as described in Figure 14a and speech information is evaluated as described in Figure 14b.

5. Discussions

In this section, we outline some design guidelines for the proposed system. The results suggest that the recognition rates using one modality or one source of information (i.e., recognition rate calculated at stage 1, represented by neuron 1 and neuron 2) are very close to other studies reported in literature which use similar modalities. However, when the outcomes of these modalities are represented as IOs and introduced to the temporal binding mechanism, the recognition rates dramatically improved. One key distinction of the proposed approach from other works is that it employs efficient processing of available information in multimodal sensors streams. The efficient processing is manifested by using a limited number of feature vectors and a limited number of elements in each vector in order to reduce the processing time of the feature vectors. For instance, the appearance-based feature vector can be constructed by applying the Gabor filter to the whole face, which may enhance the recognition rate, as calculated based on face information, and consequently increase the overall recognition performance of the system. However, the Gabor filter uses convolution operator which comes with a high computational cost. Hence, we applied the Gabor filter to selected fiducial points to reduce computational cost and exploit other biometric features in order to emphasize the real-time fashion of social human-robot interaction. The proposed approach exploits the fact that every modality participating in the encoding process of the attended subject possesses complementary information and has a discriminative level, which may be sufficient to independently identify a person and classify the individual to the correct class. In the case that the discriminative level of one modality is not sufficient to drive the system to the required threshold and finalize the identification process, it can be combined with other modalities at the intermediate level in a synergistic fashion to satisfy the required threshold, and consequently achieve higher performance.

One of the significant challenges of the person recognition tasks in social settings is that not all biometric modalities are available at the same time, due to a dynamic environment, human activities, and sensor limitations. Additionally, the nature of the HRI in social settings demands a perceptual system that is capable of providing a decision within the range of human response time; i.e., a human's reaction time. For the above reasons, exploiting the available modalities and compromising between reliability of the outcomes and fast recognition are the main characteristics of the recognition system, making it appropriate for person recognition tasks in social settings. The results show that the system achieves high performance in real-time fashion, despite the fact that not all biometric modalities are available at the same time. Table 1 shows the results of other studies that use multimodalities for person recognition tasks. Most of the reported methods use biometric modalities that are essentially invasive and require close cooperation from the attended person. Only two methods, [18] and [8], may be classified as non-invasive multimodal biometric identification systems. One shortcoming of one of these two works [8] is that the overall recognition rate is limited by the detection rates of the modalities participating in encoding the attended person. In addition, most of the works that are reported in Table 1 use one main modality as the basis to extract other auxiliary features. These are normally referred to as soft biometric features, such as gender, ethnicity, and height, which in turn are fused together in order to improve the recognition rate. Another shortcoming of all of the works reported in Table 1, including the two non-invasive approaches, is that these systems assume all modalities are available at the same times. This requirement is not normally met in real-world HRI scenarios in social settings. Thus, the main shortcoming of these approaches is that the absence of the main modality leads to failure of the overall system.

Table 1. Comparison with related works.

Approach	Biometric Modalities	Category	Accuracy	No. of Subjects
[62]	fingerprint (main) + gender, ethnicity, and height (auxiliary)	invasive	90.2%	160
[11]	face and fingerprint(main) + gender, ethnicity, and height (auxiliary)	invasive	95.5%	263
[63]	fingerprint and body weight	invasive	96.1%	62
[64]	fingerprint and iris	invasive	97.0%	21
[18]	face (main) + age and gender (auxiliary)	non-invasive	97.67%	79
[18]	fingerprint (main) + age and gender (auxiliary)	invasive	96.76%	79
[8]	skin color, hair color, eye color, weight, torso clothes color, legs clothes color, beard presence, moustache presence, glasses presence	non-invasive	not available	646
our approach	face, body, speech, and skeleton	non-invasive	100% (Figure 11c)	79

6. Conclusions

We applied, an elegant and a powerful multimodal perceptual system to address the problem of person recognition for social robots. The system can be used in a wide range of applications where a decision is expected based on the inputs from several sensors/modalities. The key distinction of this system from others is that it is non-invasive and does not require that all input stimuli are simultaneously available. The decision making process is facilitated by any modality that is rich in information and first becomes available. The system is also expected to make its decision within the same timeframe as humans (similar to duration for human response time).

In addition, the proposed system has the ability to adapt to real-world scenarios of social human-robot interactions by adjusting the threshold value which compromises between the reliability of the perception outcome and the time required to finalize the perception process. Going through the literature of person recognition systems, we note that there are almost no multimodal systems that are completely noninvasive, whereas the proposed system is noninvasive. We also note that a system that is based on "fusion" is conceptually and operationally different from the proposed architecture. The idea of fusion is to integrate the effect of several sensors with a view that each sensor by its own is not able to contribute to a correct decision; as such, the signals are fused together to enhance the decision making. The proposed system is designed based on the idea of convergence zone (as the term is used in neuroscience). This is further elaborated in Figure 1a,b. The modules "Conversion of IOs to spiking networks" and "Temporal binding" (Figure 1a) are analogous to "Multimodal Association Cortex". The process is essentially different from "fusion".

We have conducted extensive simulations and comparative studies to evaluate the performance of the proposed method. In order to generate a multimodal dataset, we combined the FERET, TIDIGITS, and RGB-D datasets to generate a new dataset that is applicable to multimodal systems. Simulation studies are promising and suggest notable advantages over related methods for person recognition.

Appendix A

L1Norm, L2 Norm, Mahalanobis distance, Cosine Similarity can be computed as (1) to (4) respectively.

$$L_1(x,y) = \sum_{i=1}^{N} |x_i - y_i| \tag{A1}$$

$$L_2(x,y) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2} \tag{A2}$$

$$Maha = \sqrt{(\overrightarrow{\mathrm{x}} - \overrightarrow{y})^{\mathrm{T}}\, \mathrm{S}^{-1} (\overrightarrow{\mathrm{x}} - \overrightarrow{y})} \tag{A3}$$

$$CosSim(x,y) = \frac{\langle x, \mathrm{y} \rangle}{||x||\ ||y||} \tag{A4}$$

where x is a feature vector represents a subject in probe set, y is a feature vector represents a subject in gallery set, S is a covariance matrix.

Table A1. Facial feature ratios.

$Ratio1 = \frac{Area\ of\ \Delta ACD}{Area\ of\ \Delta ACM_{cen}}$
$Ratio2 = \frac{Area\ of\ \Delta DHI}{Area\ of\ \Delta DJN}$
$Ratio3 = \frac{Area\ of\ \Delta JNM_{cen}}{Area\ of\ \Delta KMM_{cen}}$
$Ratio4 = \frac{Distance\ between\ point\ E\ and\ point\ G}{Distance\ between\ point\ B\ and\ point\ F} = \frac{nose\ width}{nose\ height}$
$Ratio5 = \frac{Distance\ between\ point\ A\ and\ point\ C}{Distance\ between\ point\ B\ and\ point\ F} = \frac{distance\ between\ the\ inner\_corner\ of\ the\ eyes}{nose\ height}$
$Ratio6 = \frac{Distance\ between\ point\ A\ and\ point\ C}{Distance\ between\ point\ E\ and\ point\ G} = \frac{distance\ between\ the\ inner\_corner\ of\ the\ eyes}{nose\ width}$
$Ratio7 = \frac{Distance\ between\ point\ A\ and\ point\ C}{Distance\ between\ point\ B\ and\ point\ M_{cen}} = \frac{distance\ between\ the\ inner\_corner\ of\ the\ eyes}{distance\ between\ the\ mouth\ center\ and\ the\ line\ joining\ the\ eyes}$
$Ratio8 = \frac{Distance\ between\ point\ B\ and\ point\ F}{Distance\ between\ point\ B\ and\ point\ M_{cen}} = \frac{distance\ between\ the\ nose\ tip\ and\ the\ line\ joining\ the\ eyes}{distance\ between\ the\ mouth\ center\ and\ the\ line\ joining\ the\ eyes}$

$A, B, C, D, E, F, G, H, I, J, K, L, M_c$ and N are the selected fiducial points on a face image, as shown in Figure 4a.

Appendix B

Table A2. Euclidean distance of selected skeleton segments.

(Skeleton-Based Feature)
• Euclidean distance between floor and head.
• Euclidean distance between floor and neck.
• Euclidean distance between floor and left hip.
• Euclidean distance between floor and right hip.
• Mean of Euclidean distances of floor to right hip and floor to left hip.
• Euclidean distance between neck and left shoulder.
• Euclidean distance between neck and right shoulder.
• Mean of Euclidean distances of neck to left shoulder and neck to right shoulder.
• Ratio between torso and legs.
• Euclidean distance between torso and left shoulder.
• Euclidean distance between torso and right shoulder.
• Euclidean distance between torso and mid hip.
• Euclidean distance between torso and neck.
• Euclidean distance between left hip and left knee.
• Euclidean distance between right hip and right knee.
• Euclidean distance between left knee and left foot.
• Euclidean distance between right knee and right foot.
• Left leg length.
• Right leg length.
• Euclidean distance between left shoulder and left elbow.
• Euclidean distance between right shoulder and right elbow.
• Euclidean distance between left elbow and left hand.
• Euclidean distance between right elbow and right hand.Left arm length.
• Right arm length.
• Torso length.
• Height estimate.
• Euclidean distance between hip center and right shoulder.
• Euclidean distance between hip center and left shoulder.

Table A3. geodesic distances among the projection of selected skeleton joints.

(Surface-Based Feature Vector)
• Geodesic distance between left hip and left knee.
• Geodesic distance between right hip and right knee.
• Geodesic distance between torso center and left shoulder.
• Geodesic distance between torso center and right shoulder.
• Geodesic distance between torso center and left hip.
• Geodesic distance between torso center and right hip.
• Geodesic distance between right shoulder and left shoulder.
• Geodesic distance between left hip and left knee.
• Geodesic distance between right hip and right knee.
• Geodesic distance between torso center and left shoulder.
• Geodesic distance between torso center and right shoulder.
• Geodesic distance between torso center and left hip.
• Geodesic distance between torso center and right hip.
• Geodesic distance between right shoulder and left shoulder.

Acknowledgments: This research was partially funded by NSERC and Simon Fraser University, respectively. The authors also acknowledge funds from Simon Fraser University for covering the costs to publish in open access.

Author Contributions: Mohammad Al-Qaderi is a PhD candidate undertaking his research under the supervision of Ahmad Rad.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chalabi, M. How Many People Can You Remember? 2015. Available online: https://fivethirtyeight.com/features/how-many-people-can-you-remember/ (accessed on 15 April 2016).
2. Sacks, O.W. *The Mind's Eye*, 1st ed.; Alfred A. Knopf: New York, NY, USA, 2010.
3. Brunelli, R.; Falavigna, D. Person identification using multiple cues. *IEEE Trans. Pattern Anal. Mach. Intell.* **1995**, *17*, 955–966. [CrossRef]
4. Zhou, X.; Bhanu, B. Feature fusion of side face and gait for video-based human identification. *Pattern Recognit.* **2008**, *41*, 778–795. [CrossRef]
5. Zhou, X.; Bhanu, B. Integrating face and gait for human recognition at a distance in video. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **2007**, *37*, 1119–1137. [CrossRef]
6. Palanivel, S.; Yegnanarayana, B. Multimodal person authentication using speech, face and visual speech. *Comput. Vis. Image Underst.* **2008**, *109*, 44–55. [CrossRef]
7. Gong, S.; Cristani, M.; Yan, S.; Loy, C.C. (Eds.) *Person Re-Identification*; Springer: London, UK, 2014.
8. Dantcheva, A.; Velardo, C.; D'Angelo, A.; Dugelay, J.-L. Bag of soft biometrics for person identification. *Multimed. Tools Appl.* **2011**, *51*, 739–777. [CrossRef]
9. Arigbabu, O.A.; Ahmad, S.M.S.; Adnan, W.A.W.; Yussof, S. Recent advances in facial soft biometrics. *Vis. Comput.* **2015**, *31*, 513–525. [CrossRef]
10. Feng, G.; Dong, K.; Hu, D.; Zhang, D. When Faces Are Combined with Palmprints: A Novel Biometric Fusion Strategy. In *Biometric Authentication SE-95*; Springer: Berlin/Heidelberg, Germany, 2004; Volume 3072, pp. 701–707.
11. Jain, A.; Nandakumar, K.; Lu, X.; Park, U. Integrating faces, fingerprints, and soft biometric traits for user recognition. In *Biometric Authentication*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 259–269.
12. Raghavendra, R.; Dorizzi, B.; Rao, A.; Kumar, G.H. Designing efficient fusion schemes for multimodal biometric systems using face and palmprint. *Pattern Recognit.* **2011**, *44*, 1076–1088. [CrossRef]
13. Samangooei, S.; Guo, B.; Nixon, M.S. The Use of Semantic Human Description as a Soft Biometric. In Proceedings of the 2nd IEEE International Conference on Biometrics: Theory, Applications and Systems, Arlington, VA, USA, 29 September–1 October 2008; pp. 1–7.
14. Maity, S.; Abdel-Mottaleb, M.; Asfour, S.S. Multimodal Biometrics Recognition from Facial Video via Deep Learning. *Signal Image Process. Int. J.* **2017**, *8*, 81–90.
15. Shahroudy, A.; Ng, T.-T.; Gong, Y.; Wang, G. Deep Multimodal Feature Analysis for Action Recognition in RGB+D Videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**. [CrossRef] [PubMed]
16. Frischholz, R.W.; Dieckmann, U. BiolD: A multimodal biometric identification system. *Computer (Long Beach Calif.)* **2000**, *33*, 64–68. [CrossRef]
17. Ayodeji, O.; Mumtazah, S.; Ahmad, S.; Azizun, W.; Adnan, W. Integration of multiple soft biometrics for human identification. *Pattern Recognit. Lett.* **2015**, *68*, 278–287.
18. Abreu, M.C.D.; Fairhurst, M. Enhancing Identity Prediction Using a Novel Approach to Combining Hard- and Soft-Biometric Information. *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)* **2011**, *41*, 599–607. [CrossRef]
19. Dantcheva, A.; Elia, P.; Ross, A. What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics. *IEEE Trans. Inform. Forensics Secur.* **2016**, *11*, 441–467. [CrossRef]
20. Liu, A.A.; Xu, N.; Nie, W.Z.; Su, Y.T.; Wong, Y.; Kankanhalli, M. Benchmarking a Multimodal and Multiview and Interactive Dataset for Human Action Recognition. *IEEE Trans. Cybern.* **2017**, *47*, 1781–1794. [CrossRef] [PubMed]
21. Al-Hmouz, R.; Daqrouq, K.; Morfeq, A.; Pedrycz, W. Multimodal biometrics using multiple feature representations to speaker identification system. In Proceedings of the 2015 International Conference

on Information and Communication Technology Research (ICTRC), Abu Dhabi, UAE, 17–19 May 2015; pp. 314–317.

22. Karczmarek, P.; Kiersztyn, A.; Pedrycz, W. Generalized Choquet Integral for Face Recognition. *Int. J. Fuzzy Syst.* **2017**, 1–9. [CrossRef]
23. Boucenna, S.; Cohen, D.; Meltzoff, A.N.; Gaussier, P.; Chetouani, M. Robots Learn to Recognize Individuals from Imitative Encounters with People and Avatars. *Sci. Rep.* **2016**, *6*, 19908. [CrossRef] [PubMed]
24. Asada, M.; Hosoda, K.; Kuniyoshi, Y.; Ishiguro, H.; Inui, T.; Yoshikawa, Y.; Ogino, M.; Yoshida, C. Cognitive Developmental Robotics: A Survey. *IEEE Trans. Auton. Ment. Dev.* **2009**, *1*, 12–34. [CrossRef]
25. Clemo, H.R.; Keniston, L.P.; Meredith, M.A. Structural Basis of Multisensory Processing. In *The Neural Bases of Multisensory Processes*; CRC Press: Boca Raton, FL, USA, 2011; pp. 3–14.
26. Stein, B.E. *The New Handbook of Multisensory Processes*; MIT Press: Cambridge, MA, USA, 2012.
27. Romanski, L. Convergence of Auditory, Visual, and Somatosensory Information in Ventral Prefrontal Cortex. In *The Neural Bases of Multisensory Processes*; CRC Press: Boca Raton, FL, USA, 2011; pp. 667–682.
28. Milner, A.D.; Goodale, M.A. *The Visual Brain in Action*, 2nd ed.; Oxford University Press: Oxford, NY, USA, 2006.
29. Costanzo, L.S. *Physiology*, 2nd ed.; Saunders: Philadelphia, PA, USA, 2002.
30. Halit, H.; de Haan, M.; Schyns, P.G.; Johnson, M.H. Is high-spatial frequency information used in the early stages of face detection? *Brain Res.* **2006**, *1117*, 154–161. [CrossRef] [PubMed]
31. Goffaux, V.; Hault, B.; Michel, C.; Vuong, Q.C.; Rossion, B. The respective role of low and high spatial frequencies in supporting configural and featural processing of faces. *Perception* **2005**, *34*, 77–86. [CrossRef] [PubMed]
32. Niculescu, A.; van Dijk, B.; Nijholt, A.; Limbu, D.K. Socializing with Olivia, the Youngest Robot Receptionist Outside the Lab. In *International Conference on Social Robotics*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 50–62.
33. Chellappa, R.; Wilson, C.L.; Sirohey, S. Human and machine recognition of faces: A survey. *Proc. IEEE* **1995**, *83*, 705–741. [CrossRef]
34. Kauffmann, L.; Ramanoël, S.; Peyrin, C. The neural bases of spatial frequency processing during scene perception. *Front. Integr. Neurosci.* **2014**, *8*, 37. [CrossRef] [PubMed]
35. Wallraven, C.; Schwaninger, A.; BÜlthoff, H.H. Learning from humans: Computational modeling of face recognition. *Netw. Comput. Neural Syst.* **2005**, *16*, 401–418. [CrossRef] [PubMed]
36. Baltrusaitis, T.; Robinson, P.; Morency, L. OpenFace: An open source facial behavior analysis toolkit. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016.
37. Rojas, M.M.; Masip, D.; Todorov, A.; Vitria, J. Automatic prediction of facial trait judgments: Appearance vs. structural models. *PLoS ONE* **2011**, *6*, e23323. [CrossRef] [PubMed]
38. Tien, S.-C.; Chia, T.-L.; Lu, Y. Using cross-ratios to model curve data for aircraft recognition. *Pattern Recognit. Lett.* **2003**, *24*, 2047–2060. [CrossRef]
39. Lei, G. Recognition of planar objects in 3-D space from single perspective views using cross ratio. *IEEE Trans. Robot. Autom.* **1990**, *6*, 432–437.
40. Dijkstra, E.W. A note on two problems in connexion with graphs. *Numerische Math.* **1959**, *1*, 269–271. [CrossRef]
41. Lowe, D.G. Object recognition from local scale-invariant features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; Volume 2, pp. 1150–1157.
42. Bay, H.; Tuytelaars, T.; van Gool, L. Surf: Speeded up robust features. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 404–417.
43. Shen, L.; Bai, L. A review on Gabor wavelets for face recognition. *Pattern Anal. Appl.* **2006**, *9*, 273–292. [CrossRef]
44. Serrano, Á.; de Diego, I.M.; Conde, C.; Cabello, E. Analysis of variance of Gabor filter banks parameters for optimal face recognition. *Pattern Recognit. Lett.* **2011**, *32*, 1998–2008. [CrossRef]
45. Sung, J.; Bang, S.-Y.; Choi, S. A Bayesian network classifier and hierarchical Gabor features for handwritten numeral recognition. *Pattern Recognit. Lett.* **2006**, *27*, 66–75. [CrossRef]
46. Daugman, J.G. Two-dimensional spectral analysis of cortical receptive field profiles. *Vis. Res.* **1980**, *20*, 847–856. [CrossRef]

47. Shen, L.; Bai, L. MutualBoost learning for selecting Gabor features for face recognition. *Pattern Recognit. Lett.* **2006**, *27*, 1758–1767. [CrossRef]
48. Zheng, D.; Zhao, Y.; Wang, J. Features Extraction Using a Gabor Filter Family. In Proceedings of the Sixth Lasted International Conference, Signal and Image Processing, Honolulu, HI, USA, 23–25 August 2004; pp. 139–144.
49. Serrano, Á.; de Diego, I.M.; Conde, C.; Cabello, E. Recent advances in face biometrics with Gabor wavelets: A review. *Pattern Recognit. Lett.* **2010**, *31*, 372–381. [CrossRef]
50. Miller, G.A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.* **1956**, *63*, 81–97. [CrossRef] [PubMed]
51. Kinnunen, T.; Li, H. An overview of text-independent speaker recognition: From features to supervectors. *Speech Commun.* **2010**, *52*, 12–40. [CrossRef]
52. Burkitt, A.N. A review of the integrate-and-fire neuron model: I. Homogeneous synaptic input. *Biol. Cybern.* **2006**, *95*, 1–19. [CrossRef] [PubMed]
53. Branco, T.; Häusser, M. The single dendritic branch as a fundamental functional unit in the nervous system. *Curr. Opin. Neurobiol.* **2010**, *20*, 494–502. [CrossRef] [PubMed]
54. London, M.; Häusser, M. Dendritic Computation. *Annu. Rev. Neurosci.* **2005**, *28*, 503–532. [CrossRef] [PubMed]
55. Poirazi, P.; Brannon, T.; Mel, B.W. Pyramidal neuron as two-layer neural network. *Neuron* **2003**, *37*, 989–999. [CrossRef]
56. Zhao, W.; Chellappa, R.; Phillips, P.J.; Rosenfeld, A. Face recognition: A literature survey. *ACM Comput. Surv.* **2003**, *35*, 399–458. [CrossRef]
57. Phillips, P.J.; Rizvi, S.A.; Rauss, P.J. The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1090–1104. [CrossRef]
58. Leonard, G.; Doddington, G. *TIDIGITS LDC93S10. Web Download*; Linguistic Data Consortium: Philadelphia, PA, USA, 1993.
59. Barbosa, I.B.; Cristani, M.; del Bue, A.; Bazzani, L.; Murino, V. Re-identification with RGB-D sensors. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 1–10.
60. Maass, W.; Natschlager, T.; Markram, H. A model for real-time computation in generic neural microcircuits. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 9–14 December 2002; pp. 229–236.
61. Bolle, R.M.; Connell, J.H.; Pankanti, S.; Ratha, N.K.; Senior, A.W. The relation between the ROC curve and the CMC. In Proceedings of the Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AUTO ID 2005), Buffalo, NY, USA, 17–18 October 2005; Volume 2005, pp. 15–20.
62. Jain, A.K.; Dass, S.C.; Nandakumar, K. Soft biometric traits for personal recognition systems. In *Biometric Authentication*; Springer: Berlin, Germany, 2004; pp. 731–738.
63. Ailisto, H.; Vildjiounaite, E.; Lindholm, M.; Mäkelä, S.-M.; Peltola, J. Soft biometrics—Combining body weight and fat measurements with fingerprint biometrics. *Pattern Recognit. Lett.* **2006**, *27*, 325–334. [CrossRef]
64. Zewail, R.; Elsafi, A.; Saeb, M.; Hamdy, N. Soft and hard biometrics fusion for improved identity verification. In Proceedings of the 2004 47th Midwest Symposium on Circuits and Systems, 2004 (MWSCAS '04), Hiroshima, Japan, 25–28 July 2004; Volume 1, pp. 225–228.

Appendix C.

A Two-stage Classifier Speaker Recognition System Based on Prosodic and Spectral Features via Fuzzy Inference Fusion

(under review in EURASIP Journal on Audio, Speech, and Music Processing)

A Two-stage Classifier Speaker Recognition System Based on Prosodic and Spectral Features via Fuzzy Inference Fusion

Mohammad K Al-Qaderi*, Elfituri S Lahamer, and Ahmad B Rad

Autonomous and Intelligent Systems Laboratory, School of Mechatronic Systems Engineering, Simon Fraser University, Surrey, BC, Canada

***Corresponding author: <malqader@sfu.ca>**

Abstract

We present a new architecture to address the problem of speaker identification. The proposed design integrates the prosodic features and short-term spectral features to concurrently classify a speaker's gender and his/her identity. These features are further processed by Support Vector Machine (SVM), Gaussian Mixture Model (GMM), and GMM supervector-based SVM classifiers. The outputs of these classifiers are then combined to identify the speaker. The proposed architecture works in a semi-sequential manner consisting of two stages: the first classifier exploits the prosodic features to determine the speaker's gender which in turn is used with the short-term spectral features as inputs to the second classifier system in order to identify the speaker. The second classifier system employs two types of short-term spectral features; namely Mel-Frequency Cepstral Coefficients and Gammatone Frequency Cepstral Coefficients as well as gender information as inputs to two different classifiers (GMM and GMM supervector-based SVM) which in total construct four classifiers. The outputs from the second stage classifiers; namely GMM-MFCC MLC, GMM-GFCC MLC, GMM-MFCC supervector SVM, and GMM-GFCC supervector SVM are fused at score level by the weighted Borda count approach. The weight factors are computed on the fly using Mamdani fuzzy inference system that uses the signal to noise ratio and the length of utterance as its inputs. Experimental evaluations based on TIDGITS database suggest that the proposed architecture is promising and can improve the recognition performance of the system in challenging environments where the signal to noise ratio is low and the length of utterance is short.

1 Introduction

Voice modality is central to human interactions among themselves. It is envisaged that it is also sensible that social robots communicate with humans through voice. There are three related yet different challenges that researchers are currently engaged towards realization of voice as a medium through which robots and humans interact: Who is speaking? What is being said? And which language is being spoken? This paper is concerned with the first question. Humans are better in person recognition by face rather than voice[1]. However, in machines, voice modality provides unique features in biometric person identification systems as opposed to face, fingerprint, voice, gait, or iris. The speech signal is captured dynamically over a period of few seconds; therefore, variation in speaker's model can be monitored [2]. Another advantage of using speech signal is the availability of numerous devices such as microphones, cellphone, and soundcards that can be used to capture the speech signal [3].

Speaker recognition systems can be classified into text-dependent speaker recognition and text-independent speaker recognition. The former system is employed when the speaker is cooperative and willing to pronounce fixed utterances as password or prompted by the system to pronounce pre-defined phrases that have been already registered in the system during the enrolment process. This scenario is readily used in speaker authentication systems where they extract signature(s) from a fixed utterance pronounced by an unknown speaker and verify if this utterance corresponds to the claimed identity. Even though the text-independent speaker recognition can be used as speaker verification system, the system demonstrates its advantages in other applications where the speaker needs to be identified from unconstrained utterances as in the case of identifying a speaker in social settings. Constrained utterances used in text-dependent speaker recognition system limit the functionality of the system in some applications such as social robot, speaker diarisation, intelligent answering machine with personalized caller greeting, and forensic investigation of telephone conversations. A speaker identification system for social human-robot interaction, which is the main motivation of this research, should be able to extract a voice signature from unconstrained utterances. In speaker diarisation, also known as "who spoke when", a speech signal is partitioned into homogenous segments according to speaker identity [4]. Hence, the speaker identity must be induced by extracting features from unconstrained utterances that may not have been used in the training stage.

The background speaker model is used in speaker recognition literature extensively as a way to enhance the robustness and computational efficiency of speaker recognition system [3]. Unlike other biometrics (such as face, fingerprint, iris, and hand geometry), voice biometrics is prone to substantial variability. There are many factors contributing to this variability including but not limited to changes in vocal tract and how a speaker produces speech (intrinsic-based source of variations), variation of speech signal capturing and transmission (external-based source of variations), and variation in languages and dialect spoken that are used in conversation (conversation-based source of variation). In situations referred to as within-speaker variability, the same person does not utter the same words in the same way. These conditions could arise due to physiological reasons (e.g. illness, intoxicated, and aging), emotional states (e.g. anger, happiness, and sadness), or a combination of both. In most cases, the physiological and emotional changes happen naturally and not intended to circumvent the system. Also, a person may intentionally alter his/her voice, an important factor contributing to within-speaker variability, to elude the system [5]. In addition, the performance of a speaker recognition system is affected by the technology used to capture, transmit, and save the speech signal. These scenarios are referred to as the channel variation effects, and the environmental, or background distortion. A significant research effort has been devoted to address this multifarious source of variations [6]. However, most of the research effort focuses on addressing the external-based source of variations; a considerable progress has been achieved in addressing the background noise (by using additive noise known as signal-to-noise ratio) and the channel variations (by using channel compensation techniques) [7]. Since most of the feature extraction methods used in speaker recognition systems rely on spectral characteristic of the speech signal which is highly affected by a person's health, age, and emotional state; the intrinsic-based source of variations poses huge challenge to speaker recognition systems. Among the ideal features of a speaker recognition system are low discriminant power within-speaker variability, high discriminant power between-speaker variability, robust against the aforementioned source of variations (i.e. intrinsic-, external-, and conversation- based source of variations), easy to extract, and difficult to impersonate [3]. A significant research effort has been devoted to develop feature vectors that possess some of the above characteristics [8]. In a parallel development, a large body of research has aimed on developing different speaker modeling, normalization, and adaption techniques that employ these feature vectors in order to increase the robustness of the speaker recognition system against the aforementioned source of variations [3].

Motivated by the above challenges and noting that there is no single speaker recognition model that is universally applicable in different signal-to-noise ratios and variable utterance lengths scenarios, and the absence of a “crystal ball” for speaker modeling, normalization, and adaptation; we propose a two-stage classification method for speaker identification system. The proposed system encapsulates heterogeneous classifiers that employ prosodic and short-term spectral features in a two-stage classifier in order to integrate the advantages of using different type of features and different classifiers in developing a robust speaker recognition system.

The outline of the paper is as follows: in Section 2, we review the related works and highlight the current state-of-art in speaker recognition system and the main challenges that need to be addressed. In Section 3, we present the detailed architecture of the proposed system. We will then include simulation studies and discuss the merits of the proposed architecture in Section 4. We conclude the paper in Section 5.

2 Related Works

Humans have the ability to recognize another person’s voice seamlessly without conscious effort. It is understood that various aspects of a person’s voice characteristics are implicitly and explicitly involved in the recognition process including spectral characteristics, prosody (syllable stress, intonation patterns, speaking rate and rhythm), and conversation-level features (lexicon and language). Analogous to humans, automatic speaker recognition systems employ various voice proprieties to recognize a person from his/her voice. These can be categorized into: 1) short-term spectral features; 2) voice-based and prosodic features; 3) high-level features. The short-term spectral features are computed based on short frames in the range of 10-20 *ms* and can be seen as descriptor of vocal tract characteristics. Since this category of features require a small amount of data to be extracted, they fit well with real-time applications as in the case of speaker identification in social settings [8]. Also, short-term spectral features are easy to extract and are text and language independent. Most of the automatic speaker recognition systems that have been developed in the last two decades employ short-term spectral features including MFCC, linear predictive cepstral coefficients (LPCCs), line spectral frequencies (LSFs), and perceptual linear prediction (PLP) coefficients , and Gammatone Frequency Cepstral Coefficients (GFCC), to name few [8]. However, these systems are not robust to intrinsic- and external- based source of variations and background noise. A huge research effort has been devoted to develop speaker modeling including GMM [9], GMM supervector with SVM [10], and *i*-vector system [11, 12], normalization [13, 14] and channel compensation techniques [15], and adaptation techniques [16, 17] to reduce the effect of these variations of the performance of the speaker recognition system. Impressive progress has been achieved in addressing external-based source of variations, particularly channel variations and environmental and background distortion [6]. However, the performance of the state-of-art speaker recognition systems dramatically deteriorates when short utterances are used for training/testing particularly with low SNR [18–22]. Some research studies suggest that the performance of GMM-UBM system is close to *i*-vector based system in short duration utterance and over perform it in very short utterances (less than 2 seconds) [22]. Another challenge that needs more work to address is intrinsic-based source of variations and synthesis and conversion spoofing attack; voice conversion and statistical parametric speech synthesizers may use spectral-based representation similar to the one used in speaker recognition systems that employ spectral features.

The high-level features use a speaker's lexicon (i.e. the kind of words that a speaker frequently uses in his/her conversations) to extract a signature that characterizes a speaker. Some research studies show that this category of features is robust against channel variation and background noise, but it requires substantial computational cost and difficult to extract [3, 23]. Also, this category is language and text dependent, needs a lot of training data, and is easier to impersonate. The pros and cons of prosodic features category sit in the middle of the scale between high-level feature and short-term spectral feature. Researchers suggest that the prosodic features carry less discriminant power than short-term spectral features but complementary [3]. However, due to the nature of prosody, which reflects the differences in speaking styles such as rate, rhythm, and intonation pattern. This category shows more resistance to voice synthesis and conversion spoofing attacks but it is valuable to human impersonation. One can argue that a speaker recognition system that employs short-term spectral and prosodic features is more robust than those systems that employ only one type of these features. A reasonable research effort has been devoted to fuse prosodic and spectral features in order to improve the accuracy and robustness of the recognition of a speaker age, gender, and identity [24–29]. However, most of these systems adopt either fusion at score level for prosodic-based system and spectral-based system or fusion at feature level by stacking prosodic-based feature representation with spectral-based feature representation. Fusion at score and feature level has been demonstrated in [28]; the fusion at score level was presented as a fusion of the outputs of two prosodic-based classifiers and the output of one cepstral-based classifier while fusion at feature level was performed by stacking cepstral *i*-vector with combination of the two prosodic *i*-vector representation. Some prosodic features, particularly pitch frequency F0, have demonstrated excellent performance in gender classification task [30]. Gender information can be used to enhance the GMM-based speaker recognition system in two ways. First, adaptation of speaker-dependent GMM from gender-dependent GMM-UBM is computationally efficient and demonstrates stronger coupling than adaptation of speaker-dependent GMM from gender-independent GMM-UBM [16]. Second, Reynolds et al. [31] demonstrated that increasing the population size degrades the recognition accuracy of GMM-based speaker identification system. Therefore, exploiting gender information to cluster speaker population into two groups reduces the population size and consequently improves the recognition accuracy of GMM-based speaker identification system [32]. Constructing cluster-based GMM for speaker populations improve the performance of speaker recognition system as demonstrated in [33]. Combination of prosodic features are exploited to cluster speaker populations into male and female groups to enhance the performance of emotional speech classification system [34]. Adopting gender-dependent parameterizations approach to construct GMM-based speaker recognition system improves the performance of the system, namely equal error rate and half total error rate.

The contribution of this study within this context is presentation of a novel architecture for speaker identification system that employs prosodic features as well as two types of spectral-based features (MFCC and GFCC) in order to enhance the overall recognition accuracy of the system. The system works in two stages; in the first stage, a binary classifier exploits the superiority of prosodic features to infer gender information and reduce the size of gallery set by clustering speaker population into two groups. In the second stage, the outputs of four classifiers are fused in novel way to improve the overall performance of the system; particularly in the case of short duration utterances and low SNR which is common condition in speaker identification in social settings.

3 Overview of the proposed architecture

The proposed architecture of the speaker recognition system consists of two classifiers working in a quasi-parallel fashion. The overall architecture of the system is depicted in figure 1. The upper section represents the enrollment process (training path), whereas the lower part elaborates the recognition process (testing path). The function of the feature extraction module is to transform the speaker's utterances into feature vectors that contain his/her specific characteristics. As shown in figure 1, the feature extraction module is common to both enrolment process (training) and identification process (testing). In the enrolment process, the speaker's model is trained by the feature vectors that were extracted from speech utterances by a target speaker and labeled accordingly. The recognition process is performed by extracting feature vectors from an unknown speaker's utterance which in turn is used to build the unknown speaker model. This model is subsequently matched to one of the labeled speaker models that were constructed during the enrolment process. One may infer that the feature extraction processes for both classifiers are initiated in parallel and at the same time. However, the second classifier requires gender information as well as MFCC and GFCC feature vectors in order to complete the identification process. The first classifier is a binary SVM classifier that uses prosodic feature to determine the gender of the speaker. The second classifier, which is a combination of GMM-based classifier and GMM supervector-based SVM classifiers, employs MFCCs and GFCCs feature as well as gender information to determine the identity of the speaker. In order to compute the GMM supervectors for both types of feature vectors (i.e. MFCCs and GFCCs supervectors), speaker's gender must be known. As shown in Figure 1, a binary SVM relying upon prosodic features to determine the gender of the speaker who utters a speech. The proposed speaker identification system works in two stages: First, the prosodic feature vector is used to determine if an utterance is originated from a male or a female speaker. A Binary SVM is trained using prosodic feature vector to classify the utterance into two classes (males and females). The proposed architecture incorporates the outcome of the first stage (gender classification) into the second stage where MFCCs and GFCCs feature vectors, that are extracted from the same utterance, are used to derive speaker-dependent GMM from a pre-trained gender-dependent GMM-UBM. The gender-dependent GMM-UBM is trained by utterances originated from specific gender group of speakers (male or female group). The speaker-dependent GMM derived from gender-dependent GMM-UBM shows excellent coupling to the gender-dependent GMM-UBM and requires low computational power as compared with the model derived from gender-independent GMM-UBM (i.e. the GMM-UBM is trained by utterances from male and female speakers). The resultant speaker-dependent GMMs are used to create GMM-supervectors by stacking the mean vectors of the speaker-dependent GMMs. As shown in figure 1, four classifiers have been developed by employing GMMs and GMM-supervectors. Two of these classifiers are generative-based classifier. The first one is a maximum likelihood classifier (MLC) that employs GMMs trained by MFCC feature vectors and the second classifier is a MLC that employs GMMs trained by GFCC feature vectors. The third and the fourth classifier are discriminative-based classifier; namely SVM that employs GMM-supervectors derived from GMM trained by MFCC and GFCC feature vectors respectively. Fusion at score level has been adopted to combine the outputs of all the aforementioned classifiers in a single score, namely the weighted Borda count method. In this study, the weighted Borda count method is a plausible choice as it does not require transforming all the scores of the base classifier into a common domain (i.e. no normalization process). Also, the fusion system exploits the classes ranking information of the base classifiers on the development set to compute the weights for the base classifiers (more details in section 3.3). Borda count method uses the match scores of the base classifiers to arrange the classes in descending order. Then, for each class, the Borda count is represented as the sum of the number of classes ranked below it by the respective classifier. The higher magnitude of Borda count for a class, the greater degree of agreement by the base classifiers that the test sample belongs to that class [34]. We propose a novel way to weight the Borda count for each classifier by a Mamdani fuzzy inference system [35]. Since we know that certain classifiers are more likely to outperform others at specific conditions, the weight factors can be configured to exploit the individual classifier capabilities at those conditions.

The fuzzy inference system employs the knowledge about the recognition rate trend of the aforementioned classifiers when they are evaluated on development test. This knowledge is used to derive a set of rules in the form of IF-THEN fuzzy rules based on Mamdani fuzzy inference engine in order to compute the weighting factors for all the aforementioned classifiers such that the overall recognition rate is improved. Here, we have studied the recognition rate of each classifier as a function of length of utterance and the signal-to-noise ratio (SNR). Then, for each combination of the length of utterance and the SNR, a respective rule is derived taking into consideration that each classifier is weighted by a factor proportional to its recognition rate. Also, the fuzzy rules consider the selected classifiers to be combined to complement each other from the perspectives of feature types and classifier model with priority to feature type.

In testing path, the feature extraction modules of all feature vectors used by the two classifiers are initiated at the same time. Thus, there is no noticeable delay caused by this architecture (i.e. the second classifier needs the output of the first classifier as one of its inputs in order to identify a speaker).

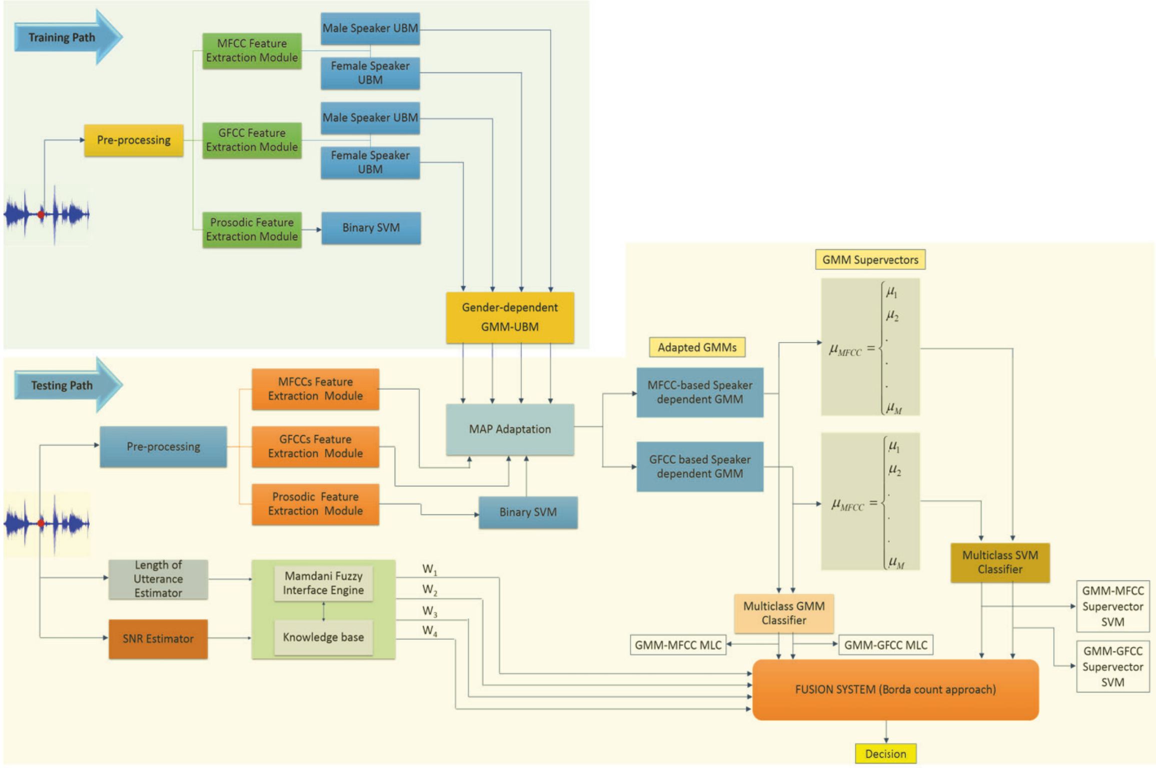

Training Path
MFCC Feature Extraction Module
Male Speaker UBM
Female Speaker UBM
Pre-processing
GFCC Feature Extraction Module
Male Speaker UBM
Female Speaker UBM
Prosodic Feature Extraction Module
Binary SVM
Gender-dependent GMM-UBM
GMM Supervectors
Adapted GMMs
$\mu_{MFCC} = \{\mu_1, \mu_2, ..., \mu_M$
Testing Path
MFCCs Feature Extraction Module
MAP Adaptation
MFCC-based Speaker dependent GMM
Pre-processing
GFCCs Feature Extraction Module
GFCC based Speaker dependent GMM
Prosodic Feature Extraction Module
Binary SVM
$\mu_{MFCC} = \{\mu_1, \mu_2, ..., \mu_M$
Multiclass SVM Classifier
Length of Utterance Estimator
Mamdani Fuzzy Interface Engine
Knowledge base
SNR Estimator
W_1
W_2
W_3
W_4
Multiclass GMM Classifier
GMM-MFCC Supervector SVM
GMM-MFCC MLC
GMM-GFCC MLC
GMM-GFCC Supervector SVM
FUSION SYSTEM (Borda count approach)
Decision

Figure 1. The architecture of the proposed speaker recognition system

3.1 Feature Extraction modules

The proposed system exploits two groups of voice-based features to identify a speaker, namely prosodic features and spectral features. The GFCC and MFCC features are adopted as spectral features. The modular representation of GFCCs and MFCCs feature extraction methods are depicted in figures 2 and 5, respectively. As shown in these figures, both MFCCs and GFCCs share the same stages except the type of the filter-banks that are applied to the resultant frequency domain signal from Fast Fourier Transform (FFT) and the compression operation. The MFCCs feature extraction method apply Mel filter-bank after FFT stage and followed by logarithmic compression and discrete cosine transform (section 3.1.1). On the other hand, in the GFCCs feature extraction method, the Gammatone filter-bank is applied to the resultant frequency domain signal from FFT before loudness compression and discrete cosine transform take place (section 3.1.2). The prosodic feature vector characterizes four areas of prosody including pitch, loudness, voice quality, and formant. Pitch and loudness information have been represented as statistical measure of fundamental frequency (F0) and energy respectively. Harmonics-to-noise ratio (HNR), jitter, and shimmer represent voice quality measurements while the first three formants characterize the fourth category of prosodic information. The complete details about prosodic feature vector are discussed in section 3.1.3.

3.1.1 MFCCs Feature Extraction Module

MFCCs feature vector is computed based on a psychologically motivated filter-bank that the spoken words cover a frequency range up to 1000 Hz. Thus, MFCCs use linearly spaced filter at low frequency below 1000 Hz and logarithmic spaced filter at high frequency above 1000 Hz. In other words, the filter-bank is condensed at the most informative part of the speech frequency (more filters with narrow bandwidths below 1000 Hz) and lengthy-spaced filter-bank is applied at other part of speech frequency as depicted in Figure 3.

As represented in Figure 2, the first step in the extraction process is to pre-emphasize the input speech signal by applying filter as:

$$Y(n) = X(n) - a * X(n-1) \qquad (1)$$

Where $Y(n)$ is the pre-emphasized speech signal , $X(n)$ is the input speech signal, and a pre-emphsized factor can be any value in the interval $[0.95, 0.98]$. In the next step (Windowing), the pre-emphsized speech signal $Y(n)$ is multiplied by smooth window function, here, we used hamming windows as:

$$W(n) = 0.54 - 0.46 * \cos\left(\frac{2\pi n}{N-1}\right), \ \ 0 \leq n < N-1 \qquad (2)$$

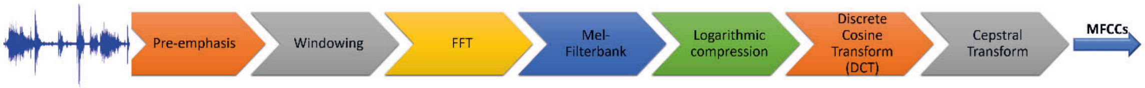


Figure 2. Block diagram of MFCCs feature extraction modules

The resultant time-domain signal is then converted to frequency domain by applying the well-known Fast Fourier Transform (FFT). The frequency range in the resultant FFT spectrum is very wide and fluctuating. Thus, the filter-bank, designed according to Mel scale, is applied in order to get the global shape of the FFT spectrum magnitude – known to contain the most distinctive information for speaker recognition. The output of the Mel-frequency filter banks is shown in Figure 4. The MFCCs are obtained by applying logarithmic compression and discrete cosine transform as in (3). The discrete cosine transform converts log Mel spectrum into time domain.

$$C_n = \sum_m^M [\log S(m)] \cos\left[\frac{\pi n}{M}\left(m - \frac{1}{2}\right)\right] \qquad (3)$$

Where $S(m), m = 1,2 \ldots., M$ is output of an M-channel filter-bank, n is the index of the cepstral coefficient. In this study, we retained the 12 lowest C_n excluding the 0th coefficient.

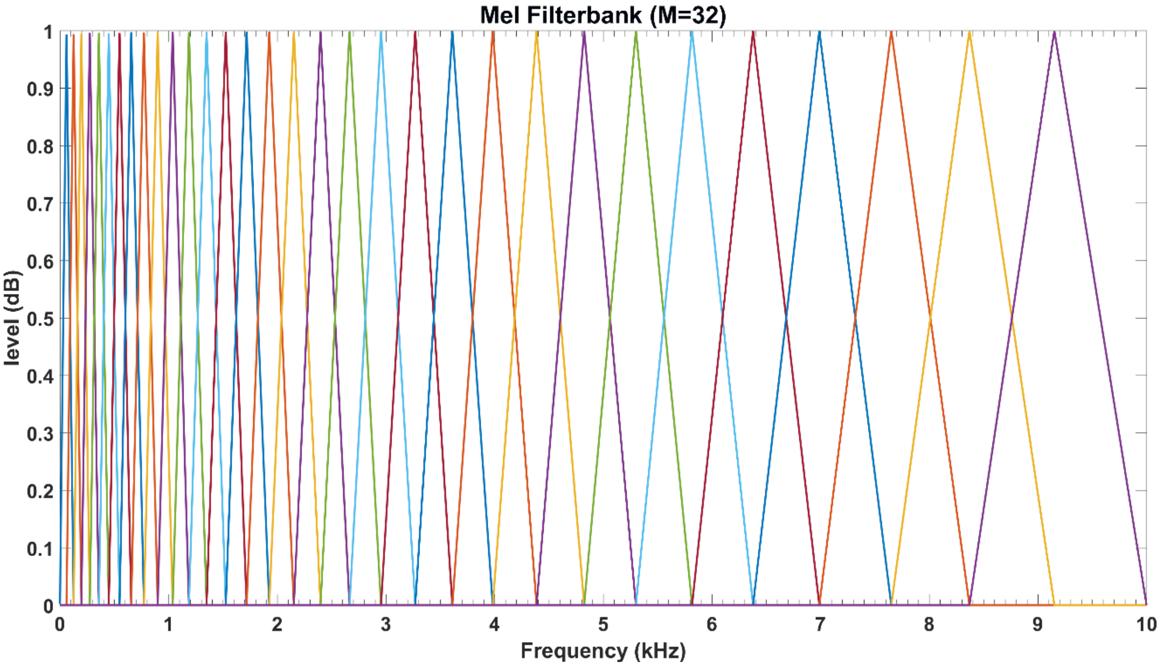


Figure 3. The triangular Mel-frequency scaled filter banks.

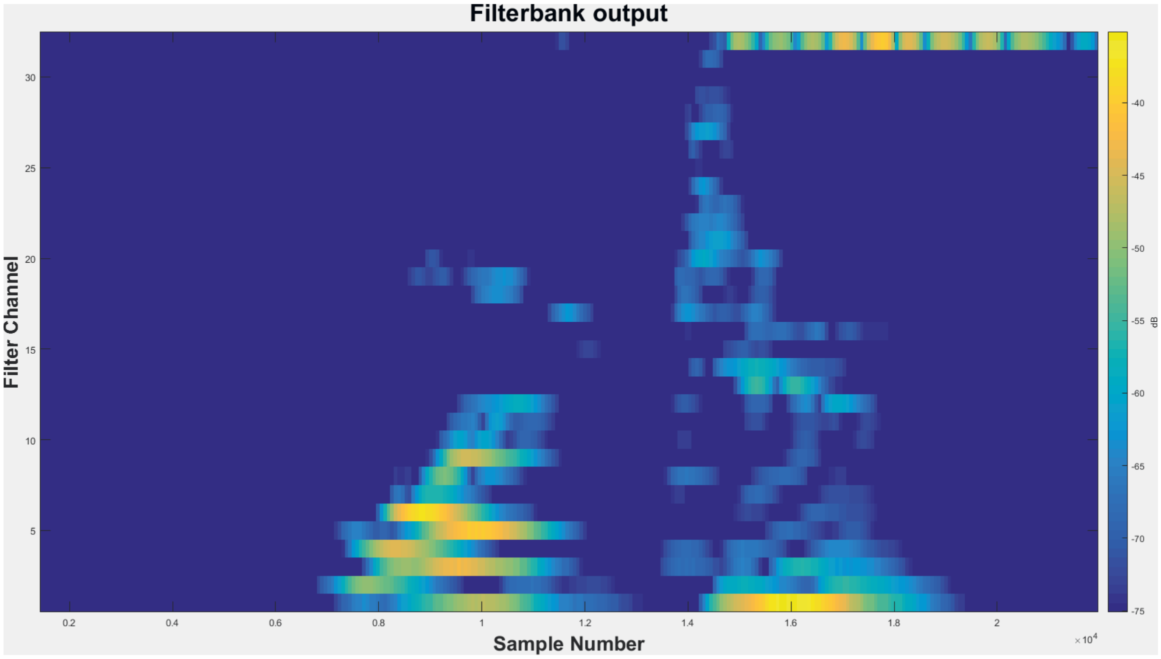


Figure 4. The output of Mel-frequency filter banks.

3.1.2 GFCCs Feature Extraction Module

The GFCCs feature vector is computed in the same way as MFCCs feature vector as shown in Figure 5. The key difference between the GFCCs and MFCCs is that the GFCCs feature extraction uses a bio-inspired Gammatone filterbank to extract the most discriminant information from FFT spectrum, which was originally designed to model the human auditory spectral response, particularly modeling the auditory processing at the cochlea. Like MFCCs, the speech signal is pre-emphasize first and followed by Windowing and FFT stages. Then, the Gammatone filterbank is applied to the resultant FFT spectrum. The impulse response of each filter in Gammatone filterbank can be represented as in (4) [36].

$$g(t) = at^{n-1}\, e^{-2\pi bt} \cos(2\pi f_c t + \varphi) \qquad (4)$$

Since a is constant, n and φ are fixed for the entire filterbank, The frequency selectivity of Gammatone filterbank is mainly defined by central frequency, f, and the filter's bandwidth b. A suggested method to compute the central frequency and the filter's bandwidth is by using an approximation to the bandwidth of human auditory filter at the cochlea. Equivalent Rectangular Bandwidth (ERB), which represents the bandwidth of series rectangular filters that used to model the human cochlea, can be used to compute the filter's bandwidth and its central frequencies. Moor [37] modeled the human auditory system using ERB as:

$$ERB(f_c) = 24.7 + 0.108\, f_c \tag{5}$$

The idea of ERB is adopted by Patterson et al. [38], to estimate the bandwidth and the center frequencies of Gammatone filter. It has been suggested that the two parameters of Gammatone filter (the bandwidth, b, and order of the filter, n) should be set as $b = 1.019\ ERB$ and $n = 4$ in order to attain a filter with good match to human auditory filter.

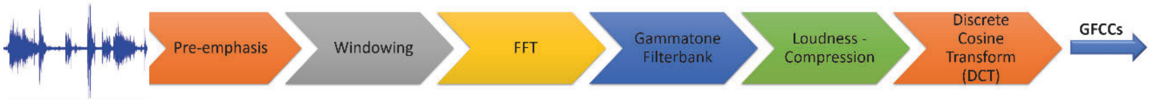


Figure 5. Block diagram of GFCCs feature extraction module.

As suggested by Moore [37], the center frequencies of the Gammatone filter are equally spaced on the ERB frequency scale. The relationship between the number of ERBs to the center frequencies, f_c, can be expressed as:

$$number\ of\ ERBs = 21.4\ log_{10}(0.00437\ f_c + 1) \qquad (6)$$

The ERB scale, which is approximately logarithmic, is defined as the number of ERBs below each frequency. This scale correlates the center frequencies with $1/_f$ distribution of frequency energy of speech signal. In other words, the frequency-dependent bandwidth of Gammatone filter produces a narrower filter at low frequencies and a broader filter at high frequencies as shown in Figure 6 (fourth-order Gammatone filterbank with 32-channel). In this study, the fourth-order Gammatone filterbank with 64-channel outputs is used to extract the GFCCs feature vectors. The GFCCs feature vectors are obtained by applying a cubic root operation (loudness-compression) and then de-correlate the feature components by applying a discrete cosine transform. Also, we retained the 22 lowest coefficients excluding 0th coefficient. The Gammatone filterbank response (termed the Cochleagram) typical spectrogram in response to a sample of speech signal from TIDIGITS are depicted in Figure 7.

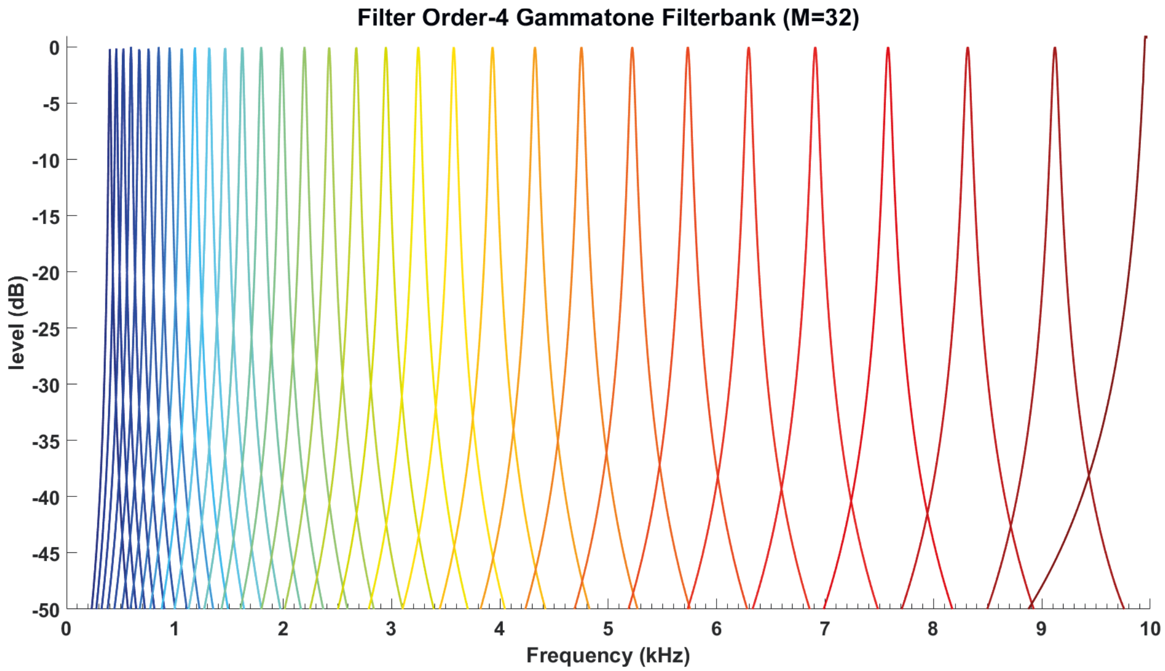


Figure 6. The frequency response of Gammatone filterbanks.

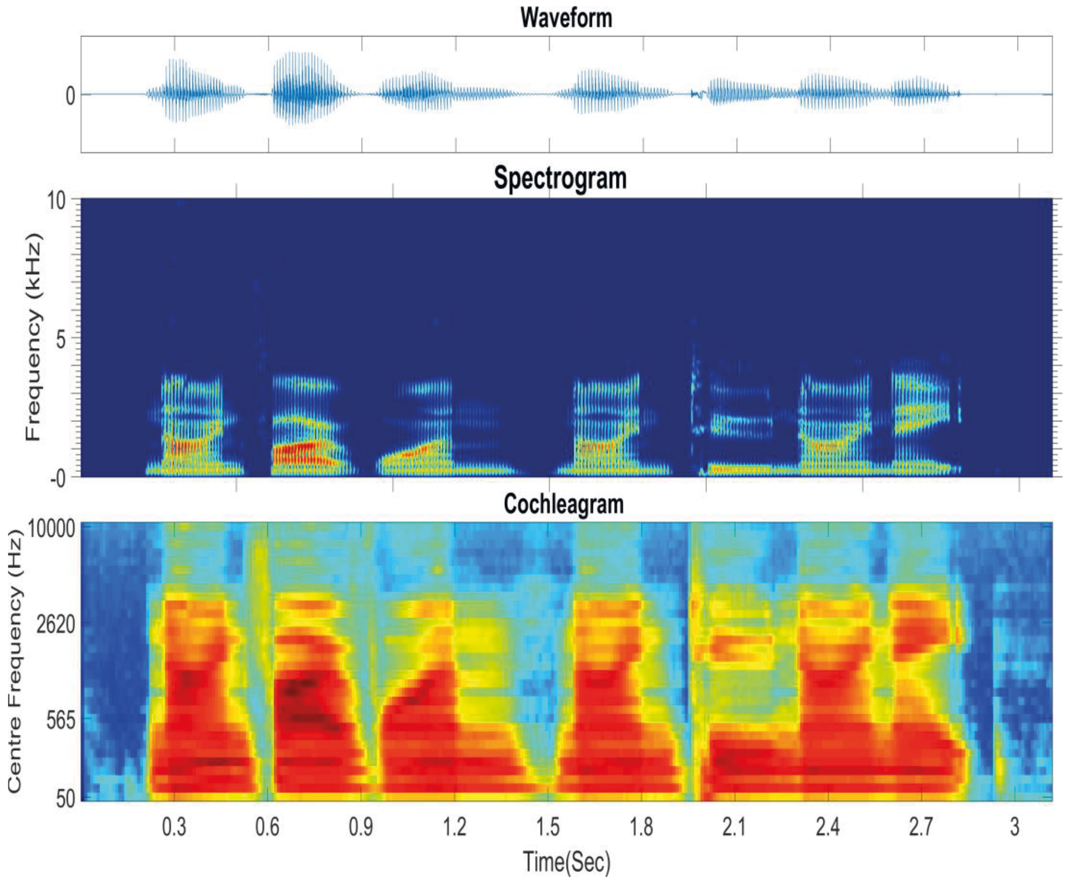


Figure 7. The Spectrogram and Cochleagram of a sample speech signal from TIDIGITS database.

3.1.3 The Prosodic Feature Extraction Module

It has been documented that 90% of speaker recognition systems have employed short-term spectral features such as MFCC, LPCC, and GFCC which have been found out to carry high power discriminant information for the speaker recognition task [8]. While the short-term spectral features span short frames of about 20-30ms and contain information correlating to timber and resonance proprieties of vocal tract, the prosodic features span longer frames of about ten to hundreds of milliseconds and correlating with non-segmental aspect of speech such as intonation, stress, rate and rhythmic organization of the speech. Since prosodic features span over long segments like syllables, words, and utterances, it is believed to contain complementary information and have more robustness for channel and background distortion. Different combinations of the prosodic parameters have been used widely for language and emotion identification, age and gender classification, and speaker authentication. The fundamental frequency (F0) is the most important parameter among these prosodic parameters. Here, selection of prosodic features from previous works of [25, 39, 40] have been adopted as feature vectors for first-stage classifier (gender classification). Particularly, various statistical measures of fundamental frequency (F0), spectral centroid (SC), spectral flatness measure (SFM), Shannon entropy (SE), harmonics-to-noise ratio (HNR), jitter and shimmer, and the first three formants have been adopted to construct one prosodic feature vector. Individually, these features are correlated with pitch accents and boundary tones, the approximate location of formants, the flatness of the spectrum, the degree of randomness of spectral probability density, the mount of noise in the speech signal, the overall periodicity of speech signal, variability of fundamental frequency, voice pathologies, and psychological characteristics of vocal tract (length, shape, and volume), respectively. HNR and the first three formant frequencies are calculated with the VoiceSauce feature extraction tool [40]. SC, SFM, and SE are computed with MIRtoolbox [41]. Absolute jitter and shimmer measurements have been extracted by using Praat voice analysis software [42]. The complete list of features and the corresponding statistical measures that used to construct the prosodic feature vector are described in Table 1 (Appendix).

3.2 Speaker Modeling (Classification algorithms)

In this section, for the sake of completeness of the paper, we consider popular classification algorithms that are widely used in speech recognition and speaker identification.

3.2.1 Gaussian Mixture Model

The well-known GMM approach has been adopted to construct the first two classifiers; namely GMM-MFCC MLC and GMM-GFCC MLC. The MFCC and GFCC feature vectors have been extracted from speech data of all speakers in training data (gallery set). For each of these feature vectors, two speaker-independent world models (a well-known universal background model (UBM)) have been created; the first UBM is trained by feature vector that extracted from female speakers and the second UBM is trained by the feature vector that is extracted from male speakers. The UBM is estimated by training M-component GMM with the standard expectation–maximization (EM) algorithm [17]. The UBM represents speaker-independent distribution of the feature vectors. Here, we use 256-

compnenent GMM to build the UBM. The UBM is represented by a GMM with 256-compnents, denoted by λ_{UBM}, that characterized by its probability density function as:

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} w_i \, p_i(\vec{x}) \qquad (7)$$

The model is estimated by a weighted linear combination of D-variate Gaussian density function $p_i(\vec{x})$, each parameterized by a mean $D \times 1$ vector, μ_i , mixing weights, which constrained by $w_i \geq 0$, $\sum_{i=1}^{M} w_i = 1$, and a $D \times D$ covariance matrix, Σ_i as:

$$p_i(\vec{x}) = \frac{1}{2\pi^{D/2}|\Sigma_i|^{1/2}} \, exp\left\{\frac{1}{2}(x-\mu_i)'({\Sigma_i}^{-1})\,(x-\mu_i)\right\} \qquad (8)$$

The training of UBM is to estimate the parameters of 256-component GMM, $\lambda_{UBM} = \{w_i, \mu_i, \Sigma_i\}_{i=1}^{M}$, from the training samples. The next step is to estimate specific GMM from UBM-GMM for each speaker in the gallery set using maximum a posteriori (MAP) estimation. The key difference between estimating the parameters of UBM and estimating specific GMM parameters for each speaker is that the UBM uses standard iterative expectation-maximization (EM) algorithm for parameter estimation. On the other hand, specific GMM parameters is estimated by adapting the well-trained parameters in the UBM to fit specific speaker model. Since the UBM represents speaker-independent distribution of the feature vectors, the adaptation approach facilitates the fast-scoring as there is a strong coupling between speaker's model and the UBM. Here, the gender-dependent UBMs have been constructed to provide stronger coupling and faster-scoring than that of gender-independent UBM. It should also be noted that all or some of GMM's parameters ($\lambda_{UBM} = \{w, \mu, \Sigma\}$ can be adapted by a Maximum A Posteriori (MAP) approach. Here, we adapted only the mean μ to represent specific speaker's model. Now, Let us assume a group of speakers $s = 1,2,3, \dots, S$ represented by GMM's

$\lambda_s = \lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_S$. The goal is to find the speaker identity $\hat{s}$ whose model has the maximum a posteriori probability for a given observation $X_k = \{x_1, \ldots, x_T\}$ (MFCC or GFCC feature vector). We calculate the posteriori probability of all observations $X_k = X_1, X_2, X_3, \ldots, X_K$ in probe set against all speakers models $\lambda_s = \lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_S$ in gallery set as (9). As $s\ and\ k$ vary from 1 to number of speakers in gallery set and number of utterances in probe set respectively, the result from (9) is $S \times K$ matrix.

$$P_r(\lambda_s|X_k) = \frac{p(X_k|\lambda_s)}{p(X_k)} P_r(\lambda_s)\Big|_{\substack{1\le s\le S\\ 1\le k\le K}} \qquad (9)$$

Assuming equal prior probabilities of a speaker, the terms $P_r(\lambda_s)$ and $p(X_k)$ are constant for all speaker , thus both term can be ignored in (9). Since each subject in probe set is represented as $X_k = \{x_1, \ldots, x_T\}$, thus by using logarithmic and assume independence between observations, calculation of posteriori probability P_r can be simplified as (10). The outputs of the two GMM-based classifiers (GMM-MFCC MLC and GMM-GFCC MLC) have been computed using (10).

$$P_r(\lambda_s|X_k) = \sum_{t=1}^{T} log\, p(x_k^t|\lambda_s)|_{\substack{1\le s\le S\\ 1\le k\le K}} \qquad (10)$$

3.2.2 GMM Supervector and Support Vector Machine

One of the challenges of exploiting information in voice modality is that the utterances are manifested with varying time duration. The dimension of feature vectors depends on the time duration of these utterances, hence the resultant feature vectors from feature extraction modules (i.e. MFCC, GFCC, prosodic) have variable dimensions. Since most of the discriminant classifiers including support vector machine (SVM) require fixed length feature vector as input, speaker recognition research community has discovered a way to represent theses time-varying utterances as a fixed-length feature vectors. The method relies upon using the parameters of speaker-dependent GMM. A speaker can be modeled as M-component GMM either by adapting a specific speaker model from UBM-GMM using MAP or by training M-component GMM with EM algorithm independently from UBM-GMM. Deriving a speaker-dependent model by adaptation approach provides a good coupling between a speaker model and UBM-GMM. Since the UBM-GMM represents a distribution of all speakers in the galley set, this coupling is desirable. Also, the adaptation approach reduces the computational cost of building speaker-dependent model and facilitate real-time response.

In this study, GMM supervector is constructed by concatenating d-dimensional mean vector of M-component speaker-dependent model that adapted from pre-trained UBM-GMM. The resultant GMM supervector with $M * d$ dimension is fed to SVM. The dimension of the MFCC-GMM supervector is 3072 ($d = 12 \: and \: M = 256$) and the dimension of the GFCC-GMM supervector is 5632 ($d = 22 \: and \: M = 256$). Principal component analysis is applied to reduce the dimension of these supervector before being fed to SVM. We refer to the two classifiers that have been trained by MFCC-GMM supervector and GFCC-GMM supervector as MFCC-GMM supervector SVM and GFCC-GMM supervector SVM, respectively.

Support vector machine is one of the most powerful discriminative classifiers with excellent generalization performance to classify any unseen data. Basically, SVM is a supervised binary classifier aims to separate the two classes by modeling a decision boundary as hyperplane; hence, adopting SVM to solve speaker verification is sensible. In speaker verification, the task is to determine if a given utterance match or does not match a target model (claimant identity). In the training stage, all training feature vectors that are extracted from the target speaker's voice samples are represented as one class and the second class is represented by all training feature vector that are extracted from the background "impostor" speaker's voice samples. SVM maps the training vector to high-dimensional space and finds an optimum hyperplane that separates the two classes (i.e. target speaker and impostor) with maximum margin. Since speaker identification is a multiclass classification problem, the well-known method One-Vs-All (OVA) SVM is adopted to extend the binary SVM to accommodate the multiclass classification task. Adopting OVA approach, which requires constructing as many binary SVM classifier as the number of classes, fits our framework of integrating the outputs of various classifiers. The output of SVM should be represented as score vector that can be interpreted either as the degree of match between a given utterance and every speaker's voice signature in gallery set or as the probability that a given utterance originates from every speaker in the gallery set. The outputs of multiclass SVM that constructed by OVA approach can be expressed as probabilistic outputs; hence, the OVA is adopted to construct multiclass SVM classifier. The probabilistic outputs are used to rank the classes and compute the Borda count value for each class.

In the proposed speaker recognition system, there is no need to unify and transform the outputs of the base classifiers to common domain. However, the outputs of the base classifiers should be expressed as scores (represent the degree of support) that are used to rank all the classes in descending order. The outputs of SVM, which is mostly expressed as a label for the predicted class that a test sample is assigned to (for example, the output of binary classifier is either +1 or -1), is not compatible with our fusion system. Therefore, the method suggested by Platt [43] is used to estimate probabilistic outputs for SVM classifier. The discriminative function of binary SVM can be expressed as (11) [44]:

$$f(x) = \sum_{i=1}^{N} \alpha_i \, y_i \, k(x, x_i) + d \qquad (11)$$

Where y_i is either $+1 \; or - 1$, and represents ideal output, x_i is support vector, d is biase, $k(x, x_i)$ is kernel function, α is weight, $\sum_{i=1}^{N} \alpha_i \, y_i = 0 \; and \; \alpha_i > 0$. The kernel function satisfies the Merce condition, so that $k(x, y)$ can be expressed as (12):

$$k(x, y) = \Phi(x)^T \Phi(y) \qquad (12)$$

Where $\Phi(x)$ is a mapping from input space (feature vector space) to high-dimensional space. Here, a radial basis function is selected as SVM kernel function and 3-fold cross validation was adopted to find best parameters for it. Mapping input feature vectors to high-dimensional space by using “kernel trick” which implicitly transforms the input vectors to high-dimensional space without explicit computation of dot product in high-dimensional space. Hence, all the dot products in SVM computations are replaced with kernel function $k(x, y)$. This implies that SVM optimizes a decision boundary (hyperplane) between the two classes in high-dimensional space without explicit computation of $\Phi(x)$. For more information about adopting SVM in speaker recognition, the reader may refer to [45], and to [44, 46] for more in-depth information about the SVM and kernel functions.

3.3 Fusion System

A large number of biometric authentication systems have adopted fusion of information at the score level to improve the overall performance of authentication systems. These systems employ various biometric modalities, different classification architectures, and feature extraction approaches. The key is not only to generate a set of feature vectors that complement each other but also to develop classifiers with satisfactory performance in diverse and challenging conditions. Here, the scores of the four classifiers (GMM-MFCC MLC, GMM-GFCC MLC, GMM-MFCC supervector SVM, and GMM-GFCC supervector SVM) are integrated using weighted Borda count method. It is worth noting that these classifiers represent both generative and discriminative approaches which intuitively can be seen as highlighting similarities and differences between classes, respectively. The weight factors are computed on the fly using a Fuzzy rule-based inference engine. The knowledge base of the fuzzy inference system is represented as IF-THEN fuzzy rules. These rules are derived by studying the recognition rate of the aforementioned classifiers as a function of SNR and the length of utterance.

In order to study the recognition behavior for individual classifiers as a function of SNR and the length of utterance, the TIDIGITS corpus [47] was divided into three equal sets; training set, development set, and testing set. The training set was used to train individual classifiers independently. All utterances in development set and testing set were distorted by three types of noise: white Gaussian noise, pink and street noise with SNR range from -5 dB to 50 dB, in increments of 5 dB. Also, all utterances in the development set were categorized into three groups (short, medium, long) based on the length of utterance. Then, the recognition rates of these trained classifiers were computed on each group and depicted as a function of SNR and rank. Each group (i.e. short, medium, long) contains utterances with range of time duration. Also, the estimation of SNR is prone to error. Thus, we fuzzified SNR and length of utterance by designing membership functions for each of them empirically as depicted in Figures 8 and 9 respectively. The SNR and length of utterance represent inputs of the fuzzy inference system and the weight factors represent the outputs of the fuzzy inference system as shown in Figure 1. The statistical measures of time duration of all utterances (i.e. min, max, mean) in each group have been used to facilitate determining the parameters of membership function of length of utterance. The SNR membership function parameters have been determined empirically based on the performance of the base classifiers on different noise levels. For each type of noise, a set of parameters has been selected relying upon the performance of the base classifiers. The knowledge base, which is represented as IF-THEN fuzzy rules, are derived from the performance of the base classifiers on the development set that categorized into three groups and distorted with three types of noise (white, pink, street). A combination of the three types of noise with three categories of length of utterance are depicted in Figure 10 to Figure 18. For each type of noise, a set of IF-THEN rules are derived relying upon the performance of the base classifiers shown in Figure 10 to Figure 18. Since the Borda count method uses ranking information to determine the winner class, the change in recognition rates of the base classifiers with respect to Rank (Rank axis in Figure 10 to Figure 18) is exploited in deriving IF-THEN rules. The rules are derived such that more weight is given to the base classifier that demonstrate big improvement in recognition rate with respect to rank axis.

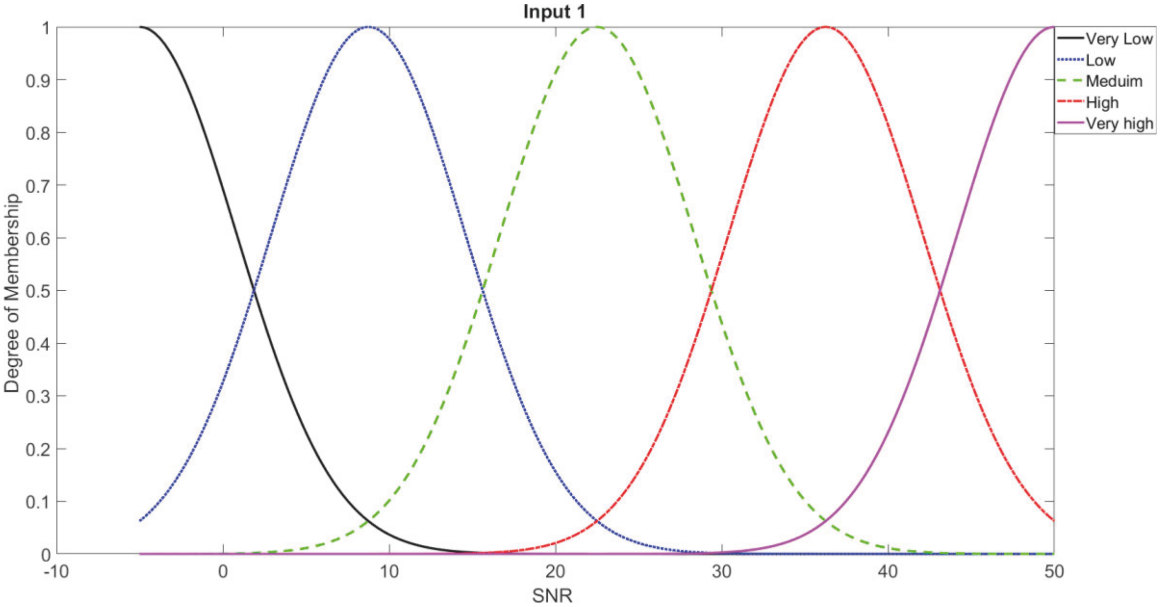


Figure 8. Fuzzy membership function for SNR input.

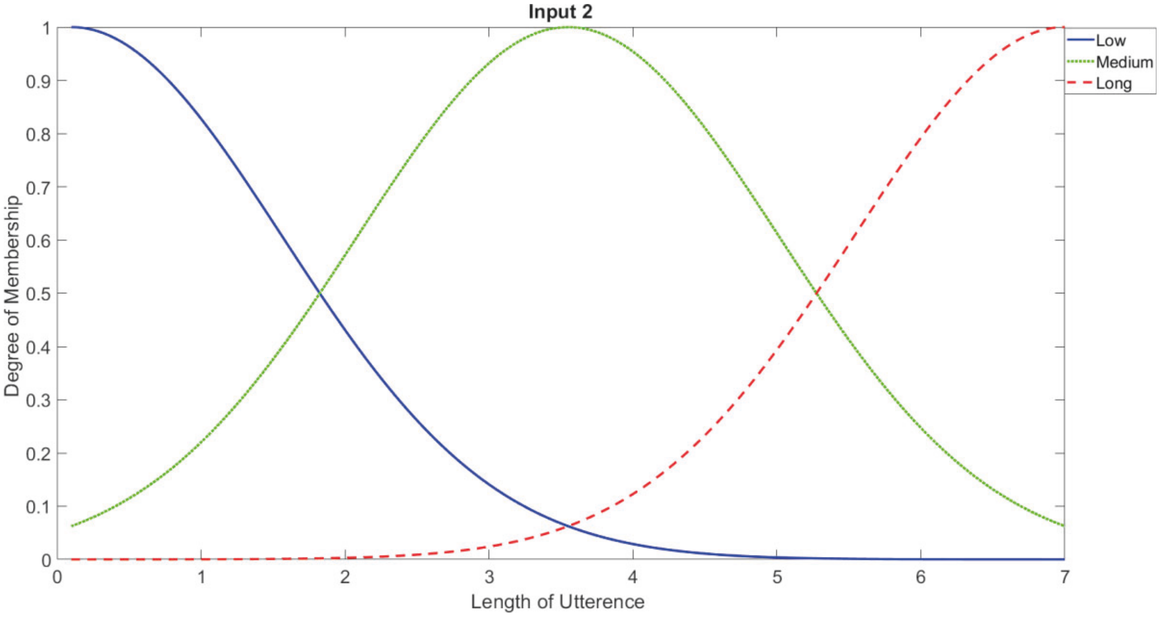


Figure 9. Fuzzy membership function for length of utterance input.

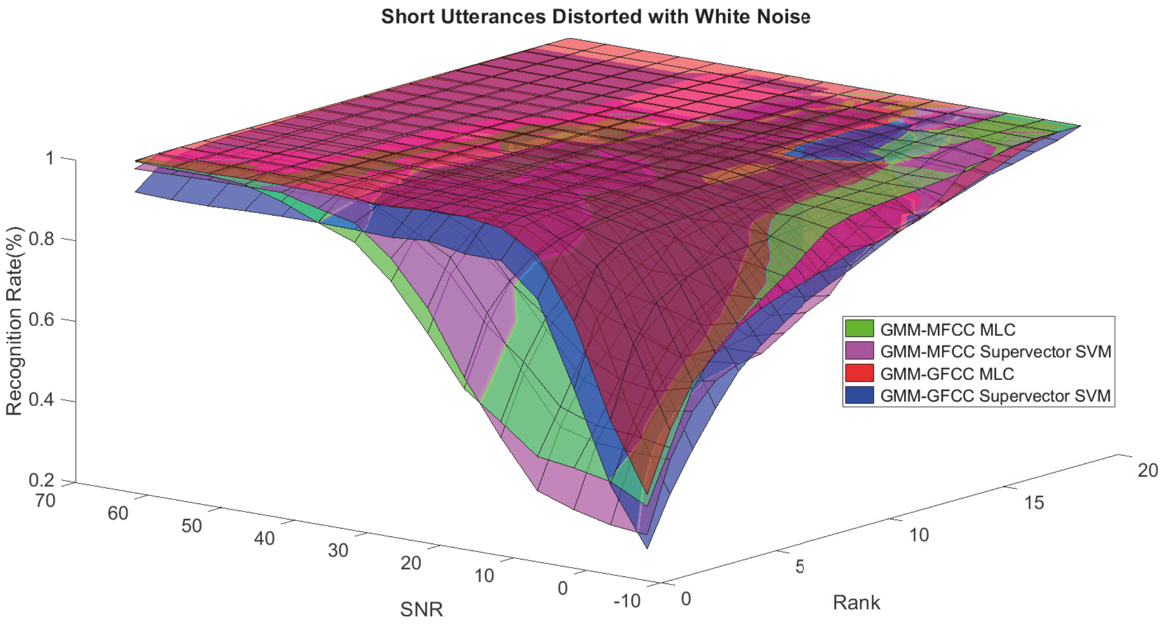


Figure 10. The 3D surface of the recognition rates for the four base classifiers on short utterances distorted with white noise.

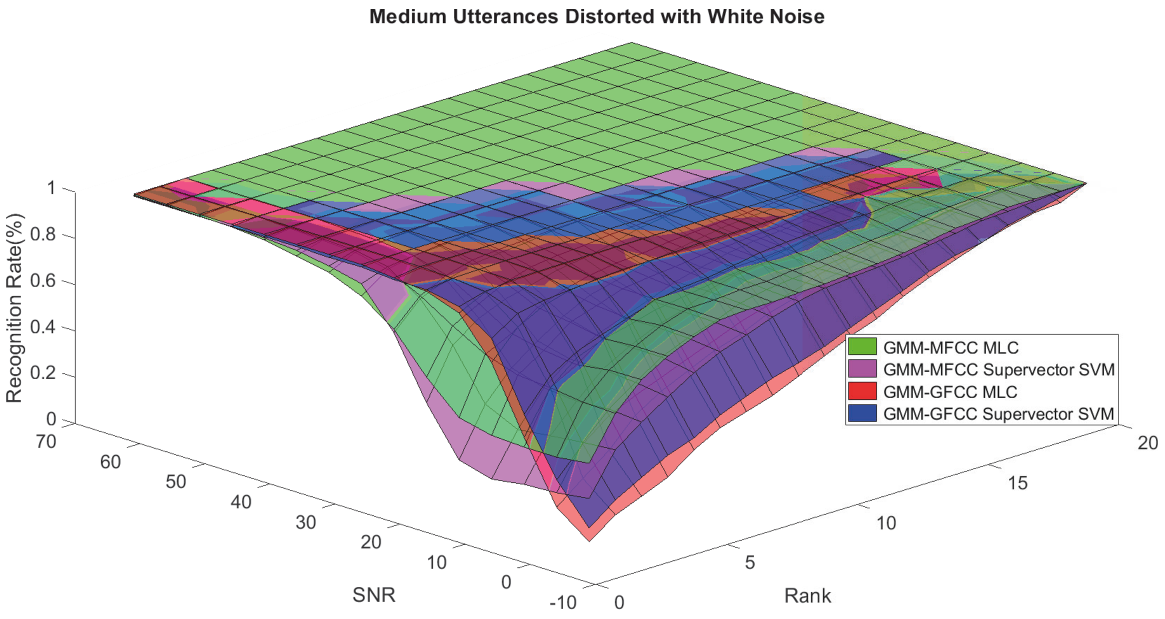


Figure 11. The 3D surface of the recognition rates for the base classifiers on medium utterances distorted with white noise.

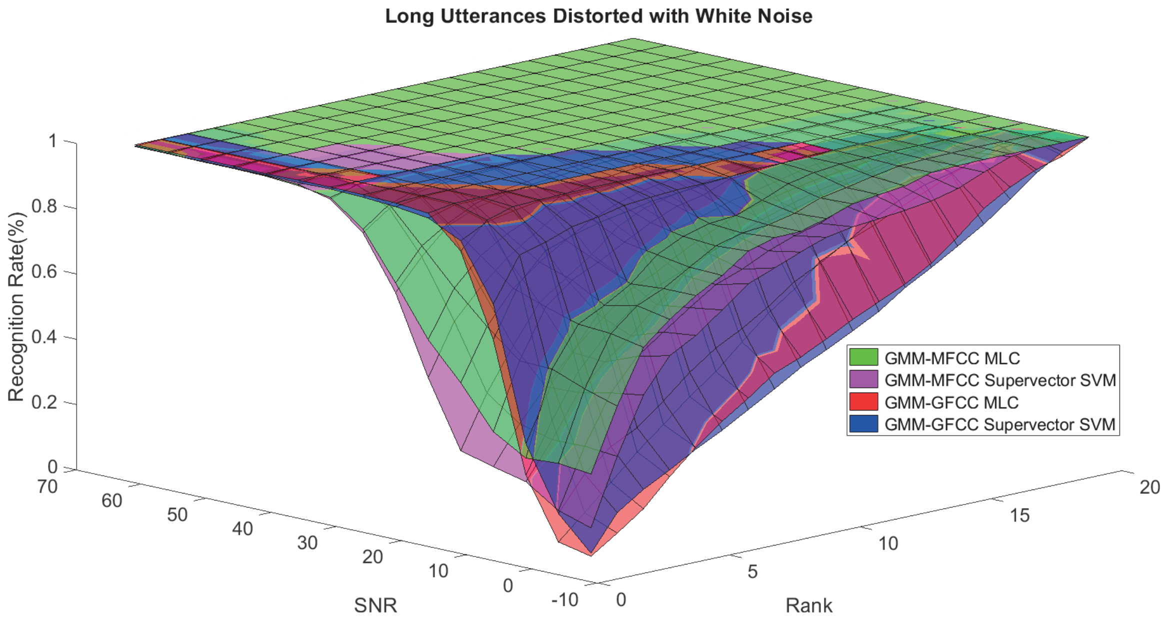


Figure 12. The 3D surface of the recognition rates for the base classifiers on long utterances distorted with white noise.

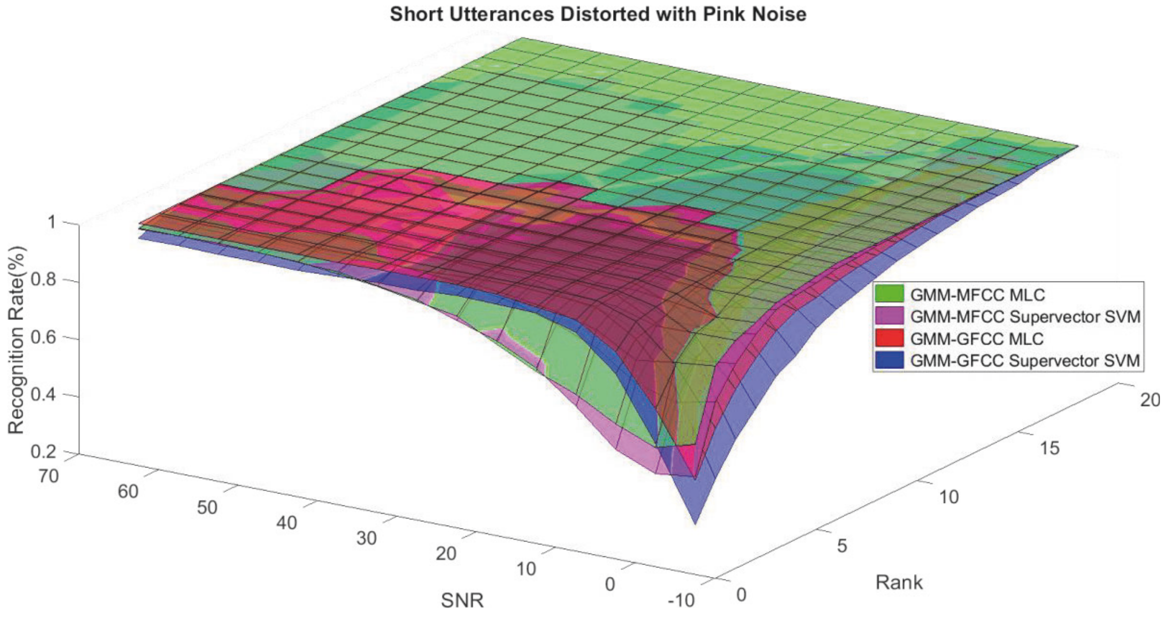


Figure 13. The 3D surface of the recognition rates for the base classifiers on short utterances distorted with pink noise.

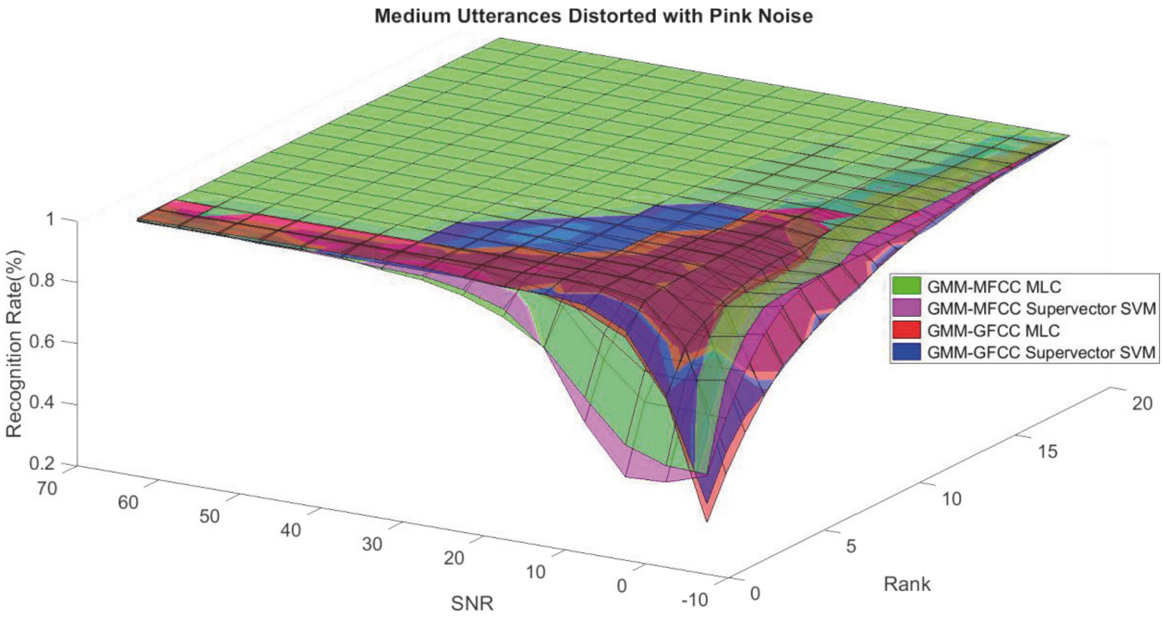


Figure14. The 3D surface of the recognition rates for the base classifiers on medium utterances distorted with pink noise.

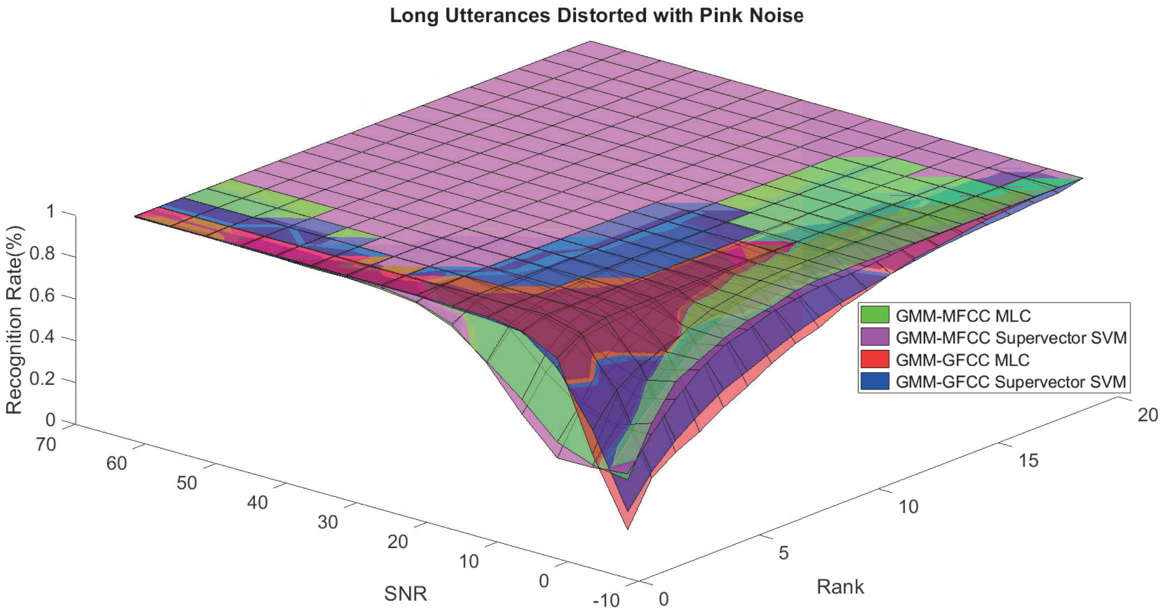


Figure 15. The 3D surface of the recognition rates for the base classifiers on long utterances distorted with pink noise.

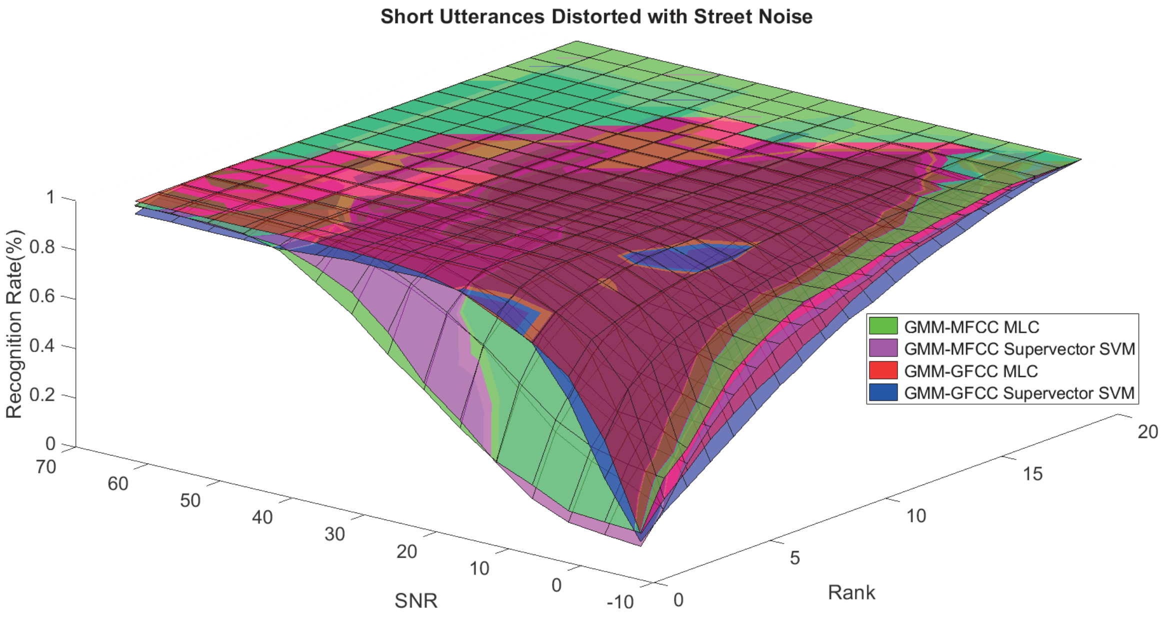


Figure 16. The 3D surface of the recognition rates for the base classifiers on short utterances distorted with street noise.

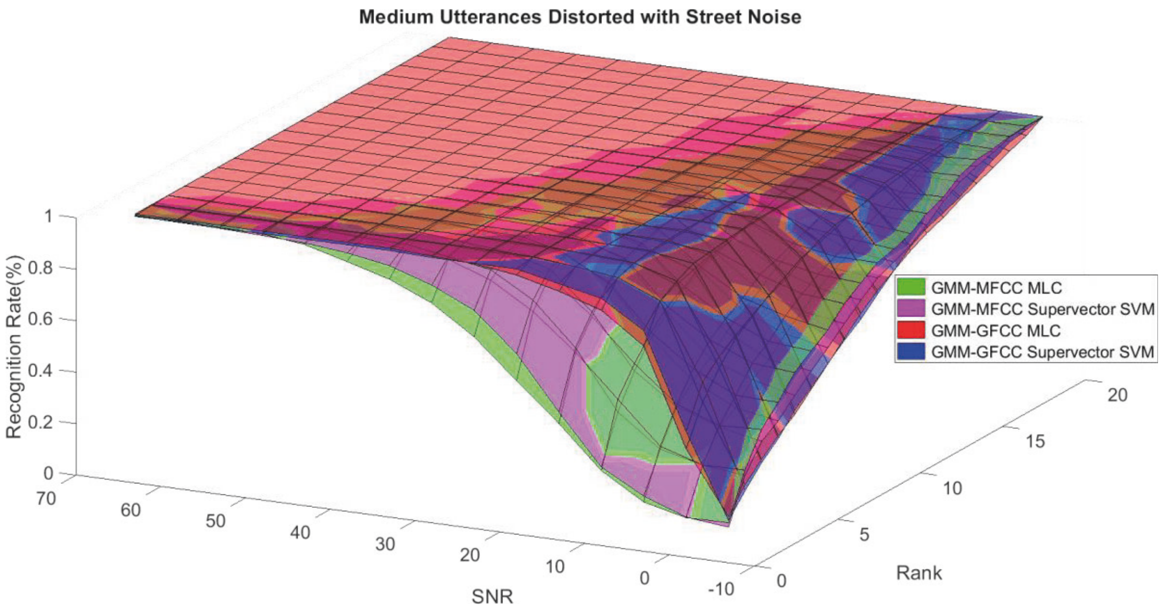


Figure 17. The 3D surface of the recognition rates for the base classifiers on medium utterances distorted with street noise.

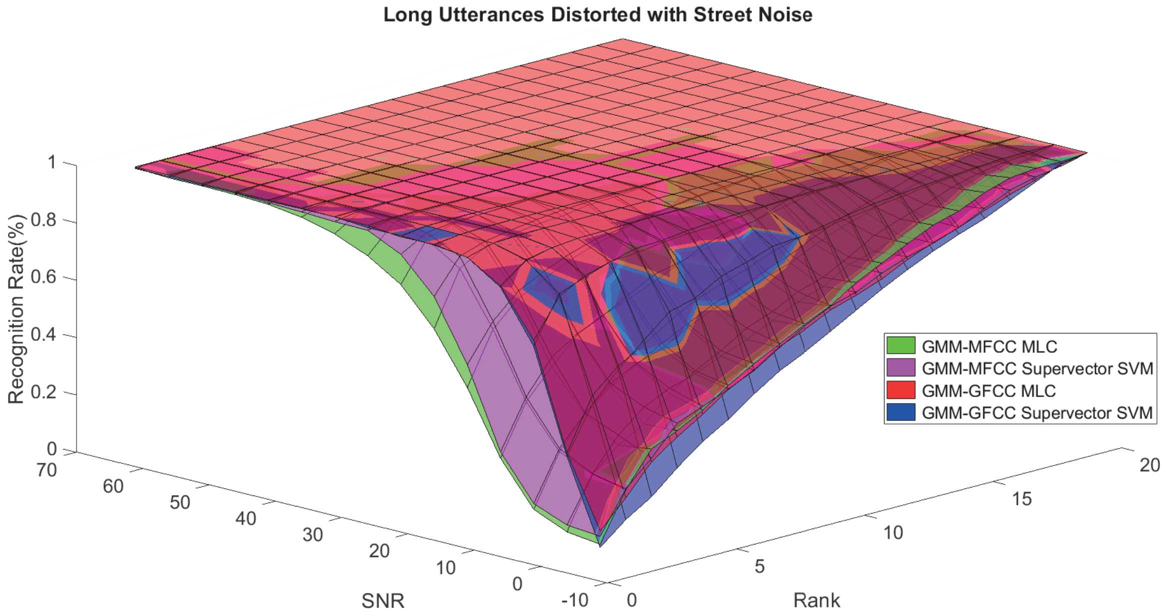


Figure 18. The 3D surface of the recognition rates for the base classifiers on long utterances distorted with street noise.

4 Experimental results

In this section, we present the experimental results to evaluate the performance of the proposed speaker recognition system. We have included three sets of simulation studies for speaker recognition identification system to demonstrate the performance of the system on three types of noise.

The TIDIGITS database is used in this research study. TIDIGITS is a speech dataset which was originally collected at Texas Instruments, Inc. The corpus was collected in a quiet environment and digitized at 20 kHz. The TIDIGITS corpus contain 326 speakers categorized into four groups (111 men, 114 women, 50 boys and 51 girls) each pronouncing 77 digit sequences. Only men and women groups were used in the experiments. Speech signals of 40 speakers (20 males and 20 females) out of 225 were randomly chosen in this study. The choice of the number of speakers was selected to facilitate comparison of our recognition system with other works. The speech samples were divided equally into three sets; namely training, development, and testing set. The training set was used for speaker modeling and training base classifiers while the development set was used to study the recognition rate as function of SNR and length of utterance, consequently derive the IF-THEN rules. The testing set was used to test the proposed systems under different noisy conditions. The testing set is comprised of unseen data, not used in the development of the system. It is worth noting that the training set is comprised of clean speech signals while the speech signals in development and training set were distorted with different noises at different SNR levels. At the test time, a prior knowledge about the type of noise is assumed. The CASA-based approach presented in [48] has been adopted to estimate the SNR of speech signal. The performance of the proposed speaker identification system was evaluated on three different noises (white, pink, street) at range of SNR (-5 dB to 65 dB). The overall recognition rates of the proposed system (fusion) and the recognition rates of the base classifiers were computed for the aforementioned types of noise. The rank-1 recognition rates of the fusion system were plotted as a function of SNR as shown in Figure 19 to Figure 21. Also, the rank-1 recognition rates of the base classifiers when they are used within the proposed system (two-stage) are depicted in the same figures. The performance of the base classifiers when they are used within the proposed system were compared with that of the same base classifiers when they are used independently without first stage classifier (i.e. without gender classification) [10, 16, 49] as shown in Figure 22 to Figure 24. To demonstrate the performance of the system in social settings, the system was tested with corpus of short utterances (0.9 to 4.5 seconds) that distorted with babble noise at range of SNR (-5 dB to 50 dB). The performance of the proposed speaker identification system was represented as rank-1 recognition rate shown in Figure 25.

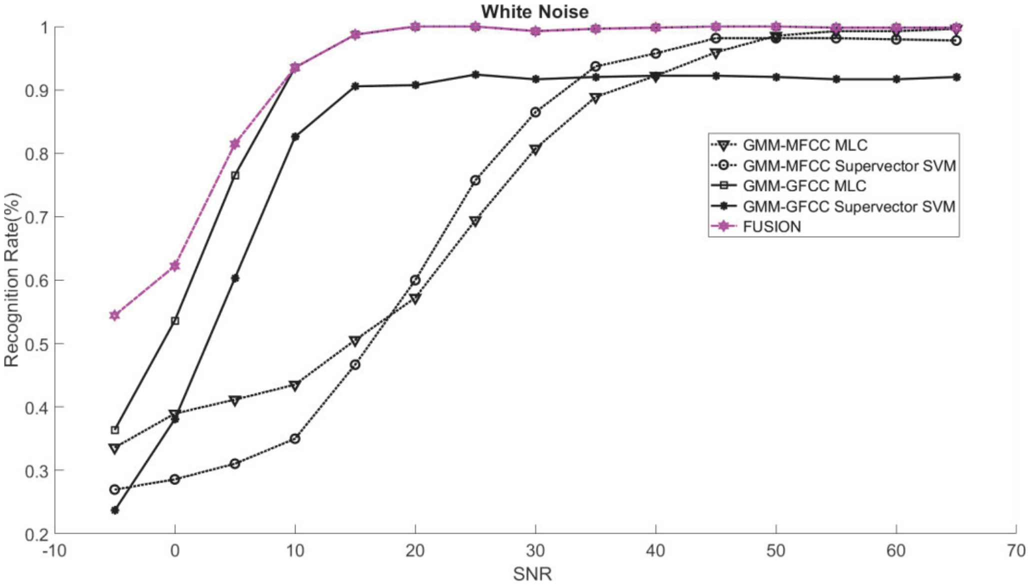


Figure 19. Rank-1 recognition rates of the fusion system and the base classifiers (two-stage) on utterances distorted with white noise.

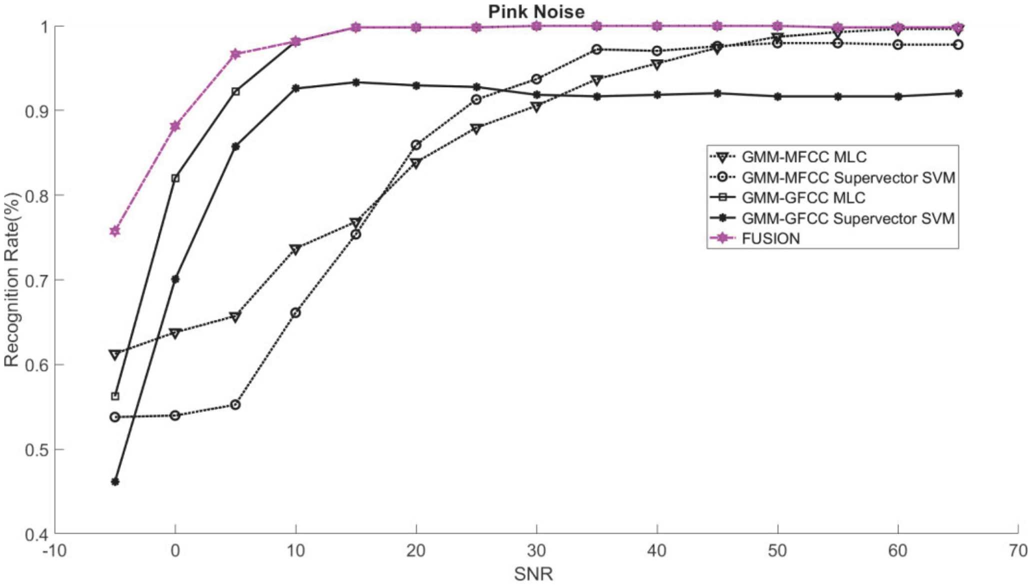


Figure 20. Rank-1 recognition rates of the fusion system and the base classifiers (two-stage) on utterances distorted with pink noise.

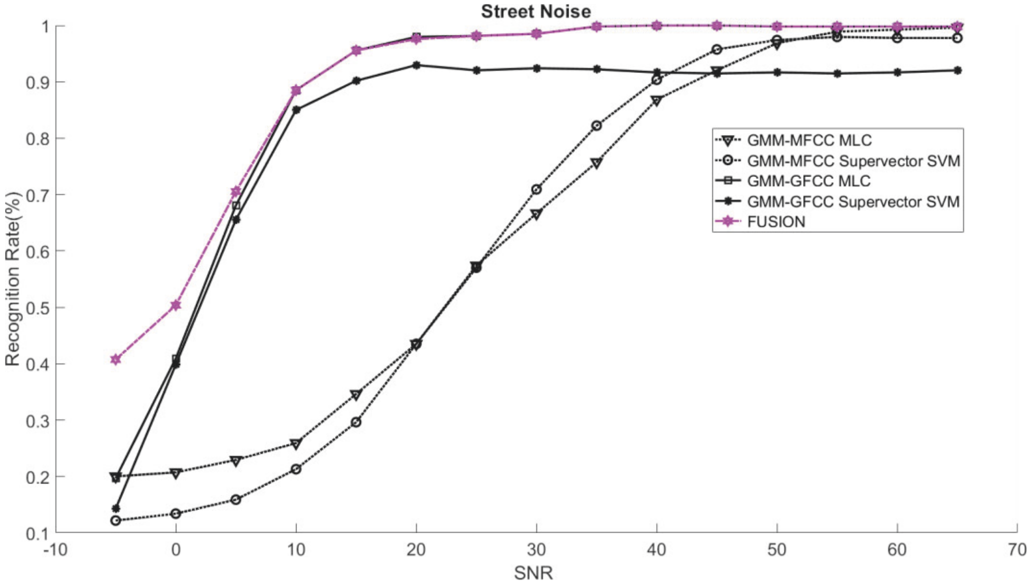


Figure 21. Rank-1 recognition rates of the fusion system and the base classifiers (two-stage) on utterances distorted with street noise.

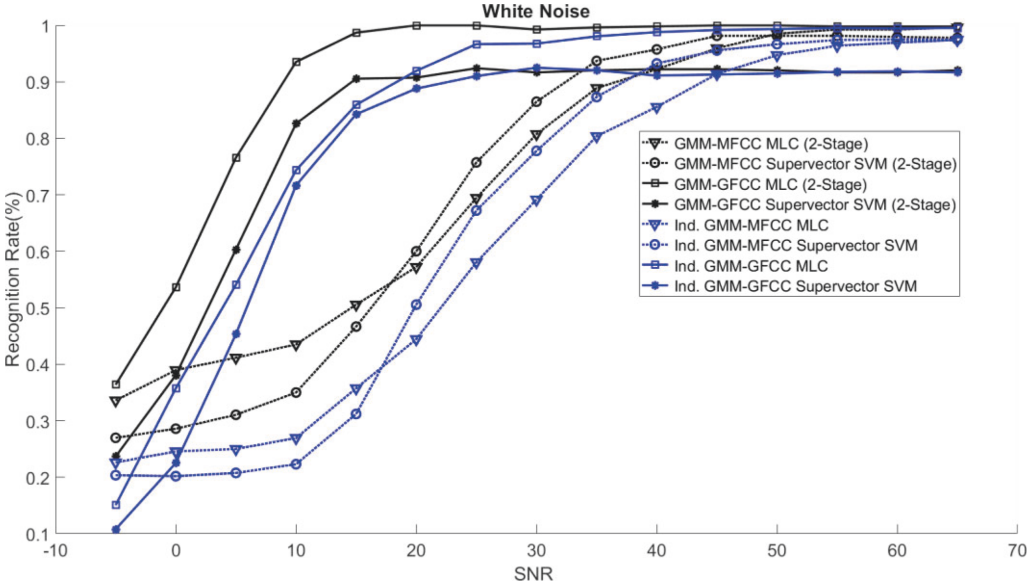


Figure 22. Rank-1 recognition rates of the base classifiers within the proposed system (black) compared with that of the same base classifiers when they are used independently (blue) on utterances distorted with white noise.

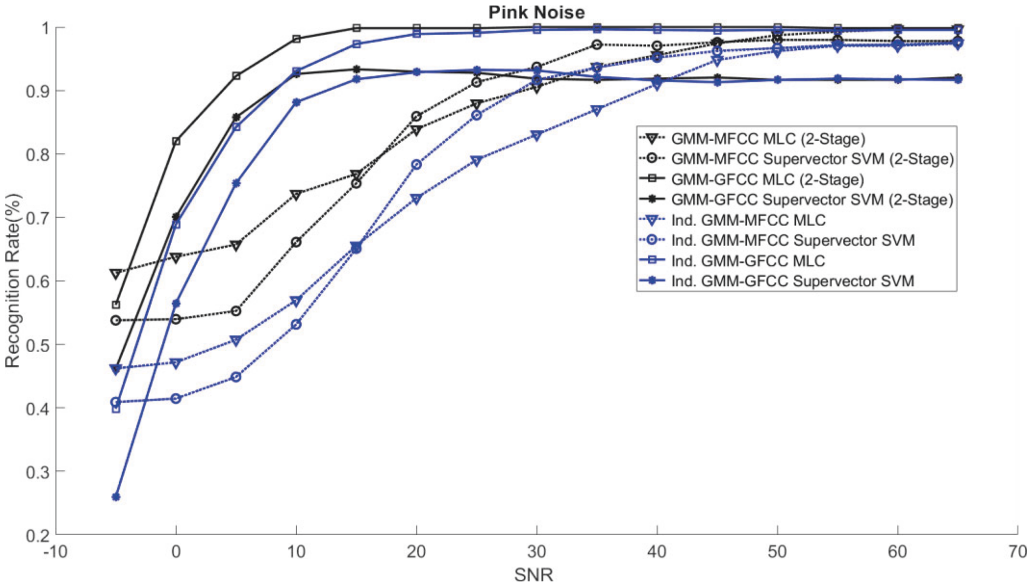


Figure 23. Rank-1 recognition rates of the base classifiers within the proposed system (black) compared with that of the same base classifiers when they are used independently (blue) on utterances distorted with pink noise.

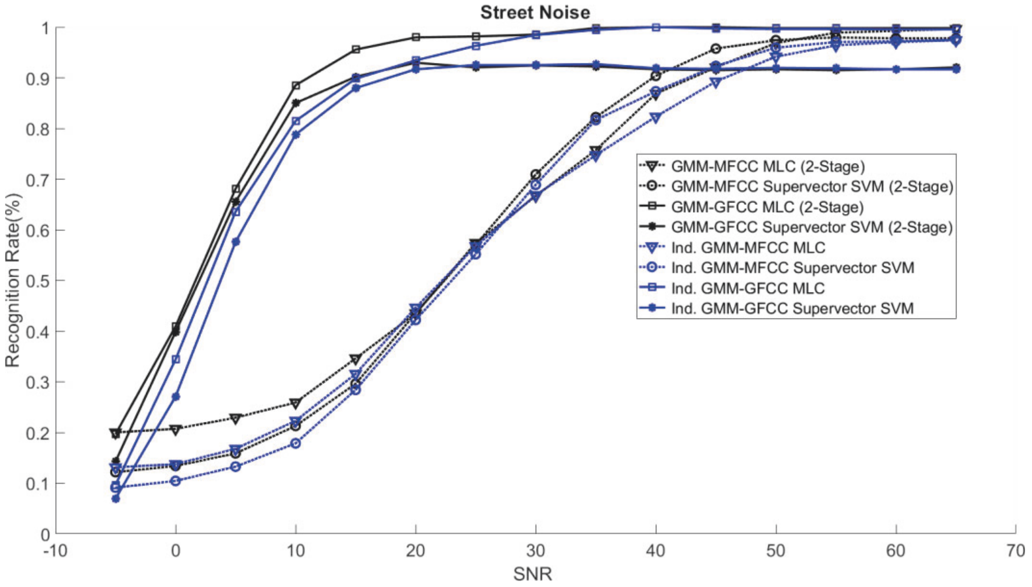


Figure 24. Rank-1 recognition rates of the base classifiers within the proposed system (black) compared with that of the same base classifiers when they are used independently (blue) on utterances distorted with street noise.

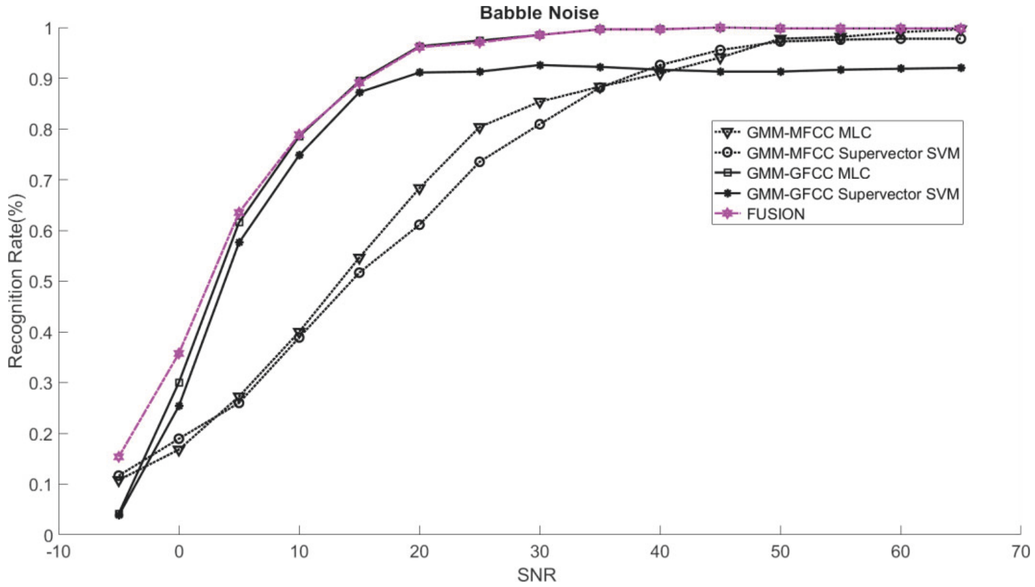


Figure 25. Rank-1 recognition rates of the fusion system and the base classifiers (two-stage) on utterances distorted with babble noise.

4.1 Discussion

The results suggest that the recognition rates of the base classifiers that are used within the proposed architecture (i.e. exploiting the gender information in the first-stage classifier) outperform that of the same classifiers when they are used independently (without the first-stage classifier). However, the recognition rates of these base classifiers and consequently the overall recognition rate of the proposed system is highly affected by the performance of the first classifier (gender classification). Also, the outcome of fusing all the base classifiers within the proposed architecture outperforms the performance of the best of the base classifiers at low SNR and match the performance of the best of the base classifiers at high SNR. The proposed fusion system exploits the knowledge about the strengths and the weaknesses of the base classifiers in order to improve the overall performance of the system. The knowledge about the strengths and the weaknesses of each base classifier at different combinations of SNR and length of utterance is used to increase the contribution of the strong classifier and to reduce the contribution of the weak classifier, consequently improve the overall performance of the system. For instance, when the speech signals are distorted with white noise, the GMM-GFCC classifier outperform all other base classifiers at low SNR and short utterance. On the other hand, at low SNR and long utterance the GMM-MFCC MLC classifier is superior to all other base classifiers. Thus, more weights are given to these classifiers when encounter similar conditions at test time. The weight factors are governed by fuzzy rules that are derived relying upon the performance of the base classifiers as discussed in section 3.3. The proposed fusion approach considers the two base classifiers that were combined such that they complement each other (i.e. the selected base classifiers need to use different features or different models given priority to the base classifier that uses different feature vectors) whenever it is possible –in the light of their performance. For instance, considering street noise, the GMM-MFCC MLC classifier was selected to be combined with GMM-GFCC MLC classifier at low to medium SNR (approximately in range of -3 dB to 12 dB) and short utterance even that GMM-GFCC supervector SVM has better performance than that of GMM-MFCC MLC classifier. However, the fuzzy inference system assigns most of the weight to GMM-GFCC MLC classifier as its performance is superior to the rest of classifiers at this specific condition.

5 Conclusions

In the absence of a unique robust speaker identification system that demonstrates superior performance for applications where the system is expected to perform in challenging scenarios such as different types of environmental noise, at different levels of environmental noise (low SNR to high SNR), and with only access to short utterances (at test time), the plausible contention is to integrate the advantages of using multi-feature speaker recognition system with multi-classifier speaker recognition system. In this study, two types of speech-based features (short-term spectral and prosodics features) and three powerful classifier systems (Support Vector Machine, Gaussian Mixture Model, and GMM supervector-based SVM classifiers) are incorporated within an elegant architecture to identify the speaker and his/her gender as by product. Exploiting prosodics features to cluster the population into two groups reduces the population size and build strong coupling between speaker-dependent model and the UBM. The reduction in population size as well as deriving speaker-dependent model from gender-dependent model, improve the recognition rates of the base classifiers. Moreover, combining the base classifiers at score level by assigning weights proportional to their performance at different conditions (combinations of SNR and length of utterance), improve the overall recognition rate of the proposed speaker recognition system particularly at low SNR and short utterance.

Appendix

Table 1 Prosodic feature vector

Name	Statistic measure
Fundamental frequency (F0)	Median, Max, and Min.
Spectral centroid (SC)	Mean and Std.
Spectral flatness measure (SFM)	Mean and Std.
Shannon entropy (SE)	Mean and Std.
Harmonics-to-noise ratio (HNR)	Mean and Std.
Jitter	Median
Shimmer	Median
The first three formant (F1,F2,F3)	Median

Declarations

a) **Availability of data and materials**

The data that support the findings of this study are available from Linguistic Data Consortium but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

b) Competing interests

The authors declare that they have no competing interests.

c) Funding

Not applicable

d) Authors' contributions

MA worked on this research project as an integral part of his PhD research. EL assisted in literature survey and data analysis. AR is the academic supervisor for this research study. All authors read and approved the final manuscript.

e) Acknowledgements

Not applicable

f) Endnotes

Not applicable

6 References:

1. Barsics, C. (2014). Person Recognition Is Easier from Faces than from Voices. *Psychologica Belgica*, *54*(3), 244–254. doi:10.5334/pb.ap

2. Rosenberg, A. E., Bimbot, F., & Parthasarathy, S. (2008). Overview of Speaker Recognition BT - Springer Handbook of Speech Processing. In J. Benesty, M. M. Sondhi, & Y. A. Huang (Eds.), (pp. 725–742). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-49127-9_36

3. Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, *52*(1), 12–40. doi:10.1016/j.specom.2009.08.009

4. Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(2), 356–370.

5. Evans, N., Kinnunen, T., Yamagishi, J., Wu, Z., Alegre, F., & Leon, P. De. (2014). Handbook of Biometric Anti-Spoofing, 125–146. doi:10.1007/978-1-4471-6524-8

6. Hansen, J. H. L., & Hasan, T. (2015). Speaker Recognition by Machines and Humans: A tutorial review. *IEEE Signal Processing Magazine*, *32*(6), 74–99. doi:10.1109/MSP.2015.2462851

7. Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Member, S., & Dumouchel, P. (2008). A Study of Interspeaker Variability in Speaker Verification, *16*(5), 980–988.

8. Tirumala, S. S., Shahamiri, S. R., Garhwal, A. S., & Wang, R. (2017). Speaker identification features extraction methods: A systematic review. *Expert Systems with Applications*, *90*, 250–271. doi:10.1016/j.eswa.2017.08.015

9. Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, *3*(1), 72–83. doi:10.1109/89.365379

10. Campbell, W. M., Sturim, D. E., & Reynolds, D. A. (2006). Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, *13*(5), 308–311. doi:10.1109/LSP.2006.870086

11. Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788–798. doi:10.1109/TASL.2010.2064307

12. Dehak, N., Kenny, P. J., Dehak, R., Glembek, O., Dumouchel, P., Burget, L., … Castaldo, F. (2009). Support vector machines and Joint Factor Analysis for speaker verification. *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 4237–4240. doi:10.1109/ICASSP.2009.4960564

13. Pelecanos, J., & Sridharan, S. (2001). Feature Warping for Robust Speaker Verification. *ODYSSEY-2001 - The Speaker Recognition Workshop*, 213–218.

14. Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *29*(2), 254–272. doi:10.1109/TASSP.1981.1163530

15. Hatch, A. O., Kajarekar, S., & Stolcke, A. (2006). Within-class covariance normalization for SVM-based speaker recognition. In *Ninth international conference on spoken language processing*.

16. Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, *10*(1–3), 19–41. doi:10.1006/dspr.1999.0361

17. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38.

18. Hanilçi, C., & Ertaş, F. (2013). Investigation of the effect of data duration and speaker gender on text-independent speaker recognition. *Computers and Electrical Engineering*, *39*(2), 441–452. doi:10.1016/j.compeleceng.2012.09.014

19. Al-Kaltakchi, M. T. S., Woo, W. L., Dlay, S. S., & Chambers, J. A. (2017). Comparison of I-vector and GMM-UBM approaches to speaker identification with timit and NIST 2008 databases in challenging environments. *25th European Signal Processing Conference, EUSIPCO 2017*, *2017–Janua*, 533–537. doi:10.23919/EUSIPCO.2017.8081264

20. Kanagasundaram, A., Vogt, R., Dean, D., Sridharan, S., & Mason, M. (2011). I-vector based speaker recognition on short utterances. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, (August), 2341–2344.

21. McLaren, M., Vogt, R., Baker, B., Sridharan, S., & Sridharan, S. (2010). Experiments in SVM-based Speaker Verification Using Short Utterances. In *Odyssey* (p. 17).

22. Poddar, A., Sahidullah, M., & Saha, G. (2015). Performance comparison of speaker recognition systems in presence of duration variability. In *2015 Annual IEEE India Conference (INDICON)* (pp. 1–6). doi:10.1109/INDICON.2015.7443464

23. Rao, K. S., & Sarkar, S. (2014). Robust Speaker Verification: A Review. In *Robust Speaker Recognition in Noisy Environments* (pp. 13–27). Springer.

24. Pandey, L., Chaudhary, K., & Hegde, R. M. (2017). Fusion of spectral and prosodic

information using combined error optimization for keyword spotting. In *2017 Twenty-third National Conference on Communications (NCC)* (pp. 1–6). IEEE.

25. Přibil, J., & Přibilová, A. (2013). Evaluation of influence of spectral and prosodic features on GMM classification of Czech and Slovak emotional speech. *Eurasip Journal on Audio, Speech, and Music Processing*, *2013*(1), 1–22. doi:10.1186/1687-4722-2013-8

26. Yücesoy, E., & Nabiyev, V. V. (2016). A new approach with score-level fusion for the classification of a speaker age and gender. *Computers & Electrical Engineering*, *53*, 29–39. doi:10.1016/j.compeleceng.2016.06.002

27. Li, M., Han, K. J., & Narayanan, S. (2013). Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Computer Speech & Language*, *27*(1), 151–167. doi:10.1016/j.csl.2012.01.008

28. Kockmann, M., Ferrer, L., Burget, L., & Černockỳ, J. (2011). iVector fusion of prosodic and cepstral features for speaker verification. In *Interspeech*.

29. Woubie, A., Luque, J., & Hernando, J. (2016). Short- and Long-Term Speech Features for Hybrid HMM-i-Vector based Speaker Diarization System. *odyssey*, 400–406.

30. Hu, Y., Wu, D., & Nucci, A. (2012). Pitch-based gender identification with two-stage classification. *Security and Communication Networks*, *5*(2), 211–225. doi:10.1002/sec.308

31. Reynolds, D. A., Zissman, M., Quatieri, T. F., O'Leary, G., & Carlson, B. A. (1995). The effects of telephone transmission degradations on speaker recognition performance. *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, *1*(x), 329–332 vol.1. doi:10.1109/ICASSP.1995.479540

32. Togneri, R., & Pullella, D. (2011). An overview of speaker identification: Accuracy and robustness issues. *IEEE Circuits and Systems Magazine*, *11*(2), 23–61. doi:10.1109/MCAS.2011.941079

33. Apsingekar, V. R., & De Leon, P. L. (2009). Speaker model clustering for efficient speaker identification in large population applications. *IEEE Transactions on Audio, Speech and Language Processing*, *17*(4), 848–853. doi:10.1109/TASL.2008.2010882

34. Mazaira-Fernandez, L. M., Álvarez-Marquina, A., & Gómez-Vilda, P. (2015). Improving Speaker Recognition by Biometric Voice Deconstruction. *Frontiers in Bioengineering and Biotechnology*.

35. Mamdani, E. H., & Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies*, *7*(1), 1–13. doi:https://doi.org/10.1016/S0020-7373(75)80002-2

36. D. Patterson, R., Nimmo-Smith, I., Holdsworth, J., & Rice, P. (1988). *An efficient auditory filterbank based on the gammatone function*.

37. Moore, B. C. J. (1997). *An introduction to the psychology of hearing / Brian C.J. Moore.* (4th ed.). San Diego .

38. Patterson, R. D., Holdsworth, J., & Allerhand., M. (1992). Auditory Models as Preprocessors for Speech Recognition. In *The Auditory Processing of Speech : From*

Sounds to Words (pp. 67–89). Berlin;New York: Mouton de Gruyter. doi:10.1515/9783110879018.67

39. Murphy, P. J. (2006). Periodicity estimation in synthesized phonation signals using cepstral rahmonic peaks. *Speech Communication*, *48*(12), 1704–1713. doi:10.1016/j.specom.2006.09.001

40. Shue, Y.-L. (2010). *The voice source in speech production: Data, analysis and models*. University of California, Los Angeles.

41. Lartillot, O., & Toiviainen, P. (2007). A Matlab toolbox for musical feature extraction from audio. In *International conference on digital audio effects* (pp. 237–244). Bordeaux.

42. Boersma, P., & van Heuven, V. (2001). Speak and unSpeak with Praat. *Glot International*, *5*(9–10), 341–347. doi:10.1097/AUD.0b013e31821473f7

43. Platt, J. (1999). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS* (pp. 61–74).

44. Cristianini, N., & Shawe-Taylor, J. (2012). *An introduction to support vector machines: and other kernel-based learning methods*. New York;Cambridge, U.K; Cambridge University Press.

45. Campbell, W. M., Sturim, D. E., Reynolds, D. A., & Member, S. (2006). Support Vector Machines Using GMM Supervectors for Speaker Verification, *13*(5), 308–311.

46. Vapnik, V. N. (2000). *The nature of statistical learning theory* (2nd ed.). New York: Springer.

47. R. Gary Leonard, G. D. (1993). TIDIGITS LDC93S10. Philadelphia: Linguistic Data Consortium.

48. Narayanan, A., & Wang, D. (2012). A CASA-based system for long-term SNR estimation. *IEEE Transactions on Audio, Speech and Language Processing*, *20*(9), 2518–2527. doi:10.1109/TASL.2012.2205242

49. Islam, M. A., Jassim, W. A., Cheok, N. S., & Zilany, M. S. A. (2016). A robust speaker identification system using the responses from a model of the auditory periphery. *PLoS ONE*, *11*(7), 1–21. doi:10.1371/journal.pone.0158520