Multimodality and the future of Personal Assistants

Stefan Dahlberg

Master's Thesis

Media Management

2015

MASTER'S THESIS	
Arcada	
Degree Programme:	Media Management
Identification number:	
Author:	Stefan Dahlberg
Title:	Multimodality and the future of Personal Assistants
Supervisor (Arcada):	Mats Nylund
Commissioned by:	Oy LM Ericsson Ab

Abstract:

The purpose of this thesis is twofold. Firstly, as it is essential to appreciate the different modalities involved in Human to Human or Human to Machine communication, the thesis starts with an overview of the basic Human and Machine communication theories before introducing the related Multimodality theory and progressing to present the Personal Assistant Model.

Secondly with regard to multimodality specifically, the thesis continues to forecast the evolution paths that Personal Assistant (PA) will likely take. By choosing the Delphi Method to research the possible PAs evolution, many interviews were conducted individually with experts and these interviews were analyzed and organized according to the method's approach. The interview outcomes are presented, which in turn help to form the conclusion of the thesis on the likely future of Personal Assistants.

The main finding is that the PAs need to be more intelligent in the future and the means to achieve more intelligence may be to incorporate multimodality in the future PAs. The conclusion of the study is that the future PAs are likely to appear in smart home environments or in help desk functions in the banking and finance sector.

Keywords:	Communication, Multimodality, Personal Assistant Model Delphi, Recommendations, Ericsson
Number of pages:	6+36
Language:	English
Date of acceptance:	

CONTENTS

Figures

Tables

FOREWORD

This Master's Thesis was written at request by Ericsson, Finland. The thesis work was started 2013 and finalized during spring 2015. I want to express my gratitude to Markku Korpi and Oy LM Ericsson Ab for giving me the opportunity to do this thesis.

I want to thank my supervisor, Adjunct Professor Mats Nylund for the excellent feedback and comments that were invaluable during the thesis writing. A big thank you to the expert panel that were able to fit the interviews in their calendars despite of their hectic work schedules Also, I want to thank my instructor at Ericsson, M.Sc. Raul Söderström, who provided me with lots of feedback and background information.

Finally, I want to thank you my family and friends, who have been supporting me and pushing me to complete this thesis work.

1 INTRODUCTION

This is a master thesis report in Media Management at Arcada, University of Applied Science in Helsinki. The supervising authority at Ericsson side is M.Sc. Raul Söderström. The supervisor at Arcada side is Ph.D. Mats Nylund. The research literature originates from within Ericsson research documentation archives, Academic journals and book sources.

1.1 Assigner

Oy LM Ericsson Ab

1.2 Problem

During the recent years the competition has significantly increased in the mobile phone industry. The fierce competition has led to a market with only a few big players, which then are in control and dominate the smart phone industry. All these major players have chosen the Personal Assistant function to be part of their standard mobile platforms. Thus, PA functions have become available for a large amount of mobile users and it has more or less become a commodity, but in which direction do the PAs evolve from here? What will the next generation PAs capabilities be? What business sectors will be PA driven in the future? What will be the environment where PAs function in the future? What are the technical obstacles hindering the PA evolving? What are the current PA flaws? These are the questions that I am trying to seek an answer to.

1.3 Purpose

This thesis exhibits that the mobile phones of today carry inputs that could be used to serve multimodal purposes. In addition, the aim of the thesis is to provide sufficient background information for the reader to be able to understand communication, modalities and multimodality concepts, which could be applied to PAs in general.

However, the main purpose of the thesis is to make a thorough investigation regarding the future evolution of PAs. Especially important is to identify trends of the future use of modalities and to evaluate the possible multimodal impacts of these modality changes. To summarize, this thesis aims to outline how the PA concept can evolve in the future and to identify some possible future PA applications that might emerge.

1.4 Research Method

The Delphi method is applied as the research method in this study. Hence, in this particular work, it means that a group of experts within Ericsson Finland are interviewed and their input is used to gather the most reliable data available in the field of PA functions. Hypothetically speaking, after an analysis of all available input data from the interviewees, it should be possible to present the best estimation for the future developments of the PA functions.

1.5 Overview of the Thesis

This thesis presents a basic and a conceptual walkthrough of the fundamental theories that are central to understanding communication, multimodality and PA functions. I believe that having all the relevant backgrounds for the subject is vital to appreciate the subsequent analysis and conclusions of the thesis. I felt that the overview of the fundamentals of human and machine communication dynamics and machine communication were lacking in all the journals and documents that I have read during researching on this topic. Thus, I made the decision to include all that material in this thesis work.

1.6 Limitations

As such, this master thesis can’t provide an in depth background and a comprehensive understanding of the subject of the communication process, multimodality and PA concepts, which would cover all angles and all aspects in these highly complex subjects.

The Delphi method was not fully applied according to its scientific definition. In this study the Delphi methodology was applied without any group sessions and expert feedback iterations. Nevertheless, an expert group was identified, which were working with-

in the PA industry at Ericsson Finland. They were all interviewed individually without knowing what their colleagues had responded. The purpose of the expert group was to anticipate and predict how the PA function would evolve in the future and to identify trends within the PA evolvement.

Revision Information

- First version was developed within Media Management learning module, which was called concept development
- Second version was developed after feedback was received on the Media Management learning module, Manage the pitch.
- Third version with a complete thesis structure applied on the subject.
- Fourth version includes changes after feedback from Mats Nylund
- Fifth version is an enhancement of the chapters and writing the abstract

2 MOBILE TRENDS

The mobile phone computing capabilities and capacity has evolved dramatically since the 1990s and its computing performance is today almost on par with a laptop. Standards like GSM, UMTS and LTE are propelling the smart phones development, which will guarantee that they further evolve to be the preferred and ultimate communication machine that connects the people across the globe and shrinks the physical dimensions of the outside world.

As I personally have been part of the mobile communication business and in a sense have been privileged to observe the mobile phone evolution through 1G to 4G, I can say that I have gained some insights of the true capabilities of the mobile phone. In my view, ever since the 1G introduction, it has been a device that is highly individual and personal. Perhaps this trend continues and the future also brings us more personalized services. Nevertheless, as the mobile phone now sits in the center of the convergence activities of various industries, it's not an easy task to forecast what will happen in the future with the smart phones and their PA functions.

2.1 Mobile Communication, Multimodality and PA aspects

As Virpi Oksman points out in her doctoral dissertation (Oksman, 2010, p.58-59), the mobile phone is a media in itself as it supports many forms of media. According to Oksman the mobile phone has demonstrated that it's capable of functioning as a mass communication device, interpersonal and personal communication device, which signifies that it's a media in itself. Oksman also brings evidence of the mobile phone of being capable to function as an enabler for TV programs as it is the device that makes the TV interactive. Also, the traditional media, such as TV, radio and newspapers are commonly viewed on a mobile phone today. The TV broadcasting trend is to move towards broadcasting in parallel over IP and terrestrial broadcasting. As the TV shows are becoming available for streaming, it means that the mobile networks can handle more capacity and the convergence of TV broadcasting and network broadcasting over mobile phones is evident according to Oksman.

To expand on the above findings of Oksman, I believe it's evident that the list of functionality that is integrated in a mobile phone gets longer and longer each day. For instance, Personal Assistants are one piece of software that has been added recently and they have already become more or less a standard feature on smartphones today. As evidence of this, I can list that Google, Apple and Microsoft deliver their own assistants on their smartphones by default. These PAs can send text messages, schedule meetings, place phone calls and understand spoken language, which is turning the PA functionality towards becoming multimodal, which means that the PAs could understand multiple inputs at the same time and possibly able to carry a meaningful dialogue at all times with the users of the PAs.

A wide range of devices such as camera, clock, calendars and other applications are common on smart phones today. Utilizing these devices properly can enable the services of a multimodal PA in a mobile phone. At the moment there are no multimodal PAs out there, since they can't handle two or more simultaneous inputs yet. For instance, a multiple input would be to use the camera input, microphone input and combine these with face recognition technology or emotional feelings recognition in the tone of voice. This could lead to a whole new range of service concept and take the PA to another level. However, this isn't available today despite that the hardware and software is in place, such as camera and microphone is present as is the PA software.

We know now that PAs are here to stay, for instance, Siri (Apple inc.), Google Now (Google) and Cortana (Windows Phone) easily demonstrates this, but how will they evolve in the future and how would technology companies possibly implement multimodal techniques in the PA feature set or will there be any multimodal PA evolution path at all?

The mobile phone was taken as a reference as it caters for multimodal inputs that are available for us and is easy to refer to. Thus, the smartphone is equipped with a range of potential multimodal inputs that could be unleashed for the PAs disposal. Who knows we could be on the verge of another revolution.

Table 1. Mobile Phone, Multimodal devices, Media Capabilities

Mobile Phone	Multimodal devices	Media Applications	Media Communications
Sony Z model	CAMERA	TWITTER	SMS (TEXTING)
	MICROPHONE	FACEBOOK	MMS
	DISPLAY	YOUTUBE	POP3, IMAP4 (E-mail)
	SPEAKERS	GOOGLE TALK	INTERNET BROWSER
		GPS, NAVIGATION, LOCATION BASED SERVICES	MPEG-4
			3GP, ALGORITHM FOR PACKING MOBILE VIDEO, AUDIO
			H.264, H.263

3 KEY CONCEPTS AND THEORETICAL FRAMEWORK

This chapter presents the fundamental theory of basic human communication, machine communication and multimodality. It also introduces the PA concept. Understanding these central concepts will positively influence the ability of appreciating the research results available in section 5.

3.1 Human communication

Vision, hearing and kinesthetic is the most dominating of human senses (Gwen van Servellen, 2009, p 24). Hence, our eyes and ears are functioning as the main organs that we use for sensing or observing our outside world and interpreting what is happening around us in the physical surroundings. For instance, our eyes and ears are receiving stimuli by light or sound. In other words, a light stimulus activates the retina in the eye and a sound stimulus activates the eardrum in our ear. The energy that is applied in a stimulus from the outside world is converted in the human sensors or organs to an internal signal, which is then passed on to the human nerve system. As the internal signal arrives to the brain an internal representation is built of the outside world in the brain. The process of applying light to the eyes retina is called a stimulus modality or sensory modality (Gwen van Servellen, 2009, p 23-26; MIAMI, 1995, p.16).

3.1.1 Modality

In accordance to the neurobiological science the modality definitions are similar to what was discussed above. In addition to vision and hearing there are several other sensory modalities. Common for all modalities or the sensory components in a human body is that they transmit nerve impulses from sensory organs to the brain. The different modalities of a human (MIAMI, 1995, p.5, p16) are illustrated in table 2 below.

Table 2. Human modalities

Sensory System	Modality	Stimulus Energy	Receptor Class	Receptor Cell Types
Visual	Vision	Light	Photoreceptor	Rods, cones
Auditory	Hearing	Sound	Mechanoreceptor	Hair cells (cochlea)
Vestibular	Balance	Gravity	Mechanoreceptor	Hair cells
Somatic senses	Touch	Pressure	Mechanoreceptor	Cutaneous mechanoreceptors
	Proprioception	Displacement	Mechanoreceptor	Muscle and joint receptors
	Temperature	Thermal	Thermo receptor	Cold and warm receptors
	Pain	Chemical, Thermal or Mechanical	Chemoreceptor, Thermo receptor, Mechanoreceptor	Polymodality Thermal, and Mechanical nociceptors
	Itch	Chemical	Chemoreceptor	Chemical nociceptors
Gustatory	Taste	Chemical	Chemoreceptor	Taste buds
Olfactory	Smell	Chemical	Chemoreceptor	Olfactory sensory neurons

3.1.2 Perception and cognition

Perception is the decoding or interpreting of the raw data that is received to our brain via transduction. Transduction is the process, which is referring to the raw data conversion from the receptors to an electrochemical pattern that the neurons in our brains understand. This process, when the neurons in our brain, as a part of our central nervous system, arranging the raw data that arrives from the receptors and trying to make a higher level abstraction of the physical input is called perception.

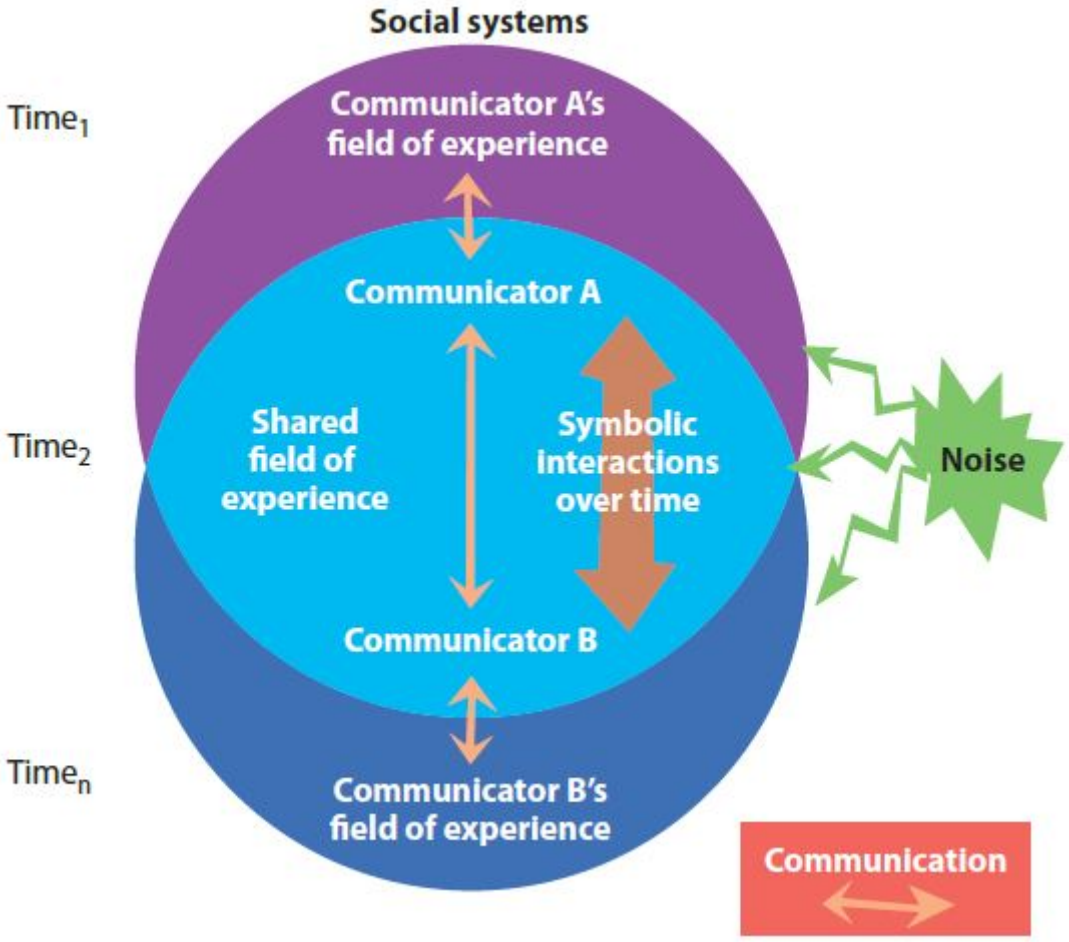
Cognition again, is defined to be the ability to learn, reason, and to understand and produce language. Thus, cognition is an important part of what is defined in psychology as a mind. Furthermore, a mind is an entity that can generate emotions and behavior out of

the information received through the sensors. In daily life we are affected by our cognition to make decisions, for instance waiting and making the decision to board the bus number 63, which takes us to our preferred destination in the morning.

3.1.3 The communication process

There are several models that describe the communication process, namely the linear, the interactive and the transactional model. In general, the communication process is defined as being the way for humans to convey their thoughts, ideas and feelings to other human's attention. In fact, communication is an essential part of our daily life and throughout the entire life. The key elements in any communication activity, where information is send and received, are the encoding phase, transmission phase, decoding phase and the feedback phase. In the linear communication flow the message is passed from the source to the receiver in one direction only. Hence, the linear model is without a feedback loop. In the interactive model the feedback loop is incorporated and there is an interchange of the information between the source and the receiver. Observe that the transmission medium at some point of the decoding and encoding will be using the human senses to make it understandable and meaningful for the brain. The transmission medium or channel can be, for instance, verbal or nonverbal communication. In other words, if there is a dialogue between two persons and they are having a conversation, it means that both parties are providing feedback to each other continuously, which makes the communication interactive or defined as being circular communication (Gwen van Servellen, 2009, p 40). The term circular is referring to the continuous feedback loop that is established between both parties interacting and influencing each other at all times. According to TJ Wood (Wood, J. 2011, Communication in our lives, Ch. 1) this interactive model isn't complete without taken into consideration also the time dimension, where two persons knowing each other for long time can speak more openly than if they just met. JT Wood also emphasizes that the interaction model is flawed as it labels the speakers as receivers and senders, while it is more correct to understand that these roles are not fixed and that the roles of receiver and sender constantly changes during a discourse. JT Wood is referring to the transactional model to be used as the preferred model as it's the most evolved and it depicts the communication experience as being mutually interactive.

Figure 1. Transactional communication model (JT Wood, Communication in our lives, 6th Edition, 2012, 2009, 2006)



3.1.4 Communication types

The basic types of human communication are defined as verbal and nonverbal communication. The verbal communication is strongly tied to the language and therefore it refers to using words to pass the message to the receiver. The verbal communication can also be written communication using words. The nonverbal communication methods (Mark Butland, An Introduction to Human Communication, Ch. 4) and the verbal communication methods are listed in table 3.

Table 3. Verbal and Nonverbal methods of communication

Verbal Communication (Oral, Spoken word)	**Verbal Communication (Written)**	**Nonverbal Communication (Messaging without words)**
Face to Face dialogue	Contracts, Laws	Kinesics - Body conveys the message • Gestures • Posture • Facial expressions
Telephone	SMS, Text messaging	Occulesics – Eye Movements • Open a communication channel • Demonstrate concern • Gather feedback • Moderate anxiety
Video Conferences	Books	Haptics – Touch conveys the message • Emotions • Sympathy • Care
Lectures	E-mail	Proxemics – Spacial Distance • Intimate distance • Personal distance • Social distance • Public distance
Voice messages	Documents	Vocalics – sighs (aahh, ummm) • Pitch • Rate • Volume • Tempo • Accents
Conference Events	Chat	Chronemics – Use of time • Punctuality • Willingness to wait
		Appearance • Choice of clothing color • Hairstyle • Tattoo • Body shape (thin / thick)

3.1.5 Communication levels

The communication levels are divided into intrapersonal, interpersonal and mass communication. The intrapersonal communication is the dialogue with yourself, for instance, when you are keeping a diary. Another type of intrapersonal communication would be to think, which could be interpreted as a mental conversation. The interpersonal communication is about having a conversation between exactly 2 people. Mass communication types would be lectures, public speeches, where there is an audience as a receiving party.

3.2 Machine to human communication

There are several inbuilt sensors in a mobile phone. On a general level a sensor is perceived to be a device that measures a physical quantity and converts the measured input data to observable output data for an observer. As an example, the mobile phone has a location data sensor that collects input information from GPS signals and produces an output of its location, which is then displayed on the screen of the mobile device and the location can be understood by a mobile phone user. If considering human capabilities, these mobile sensors can be seen as expanding on the human capabilities as the mobile devices in a way enable more sensory information for the humans to pick up from their environment surroundings.

3.2.1 Sensors and devices

The common mobile sensors that are embedded in a smart phone consist of location, motion, environmental and position sensors. For instance, the location sensors would provide a longitude and latitude output for a human to read visually using a mobile application and the mobile screen or alternatively put your location directly on a map, which displays exactly where your physical location is at present. In Figure 5 the dominant mobile platforms and their sensors are showed and they can be used by mobile applications if required or allowed by the mobile user profile. All types of sensors are potentially enablers of a richer machine to human communication, in case these properties can be taken into use by an interface that can take advantage of these properties and apply them between man and machine communication.

Table 4. Sensor availability, Android platform 4, Apple iOS and Windows

Sensor Types	Android 4.0 Google	iOS5 by Apple	Windows 8 by Microsoft
ACCELEROMETER	YES	YES	YES
TEMPERATURE	YES	YES	YES
GRAVITY	YES	YES	YES
GYROSCOPE	YES	YES	YES
LIGHT	YES	YES	YES
MAGNETIC_FIELD	YES	YES	YES
ORIENTATION	YES	YES	YES
PRESSURE	YES	NO	NO
PROXIMITY	YES	YES	YES
HUMIDITY	YES	NO	NO

3.3 Multimodality

The term multimodality is usually referring to either the human or machine input/output channels. Due to this it's important to understand the human side fundamentals and what is meant by communication, human senses and cognition, since these do form the fundamental modeling entity that can be defined as a front end to a multimodal interface, namely the human. Understanding the system side, which is describing the computer input devices and sensors is equally important. The computer side can be described as the back end entity of a multimodal interface. The ultimate goal is to match entities on the both sides with a multimodal interface that is able to produce an optimal fit between these entities, man and the machine.

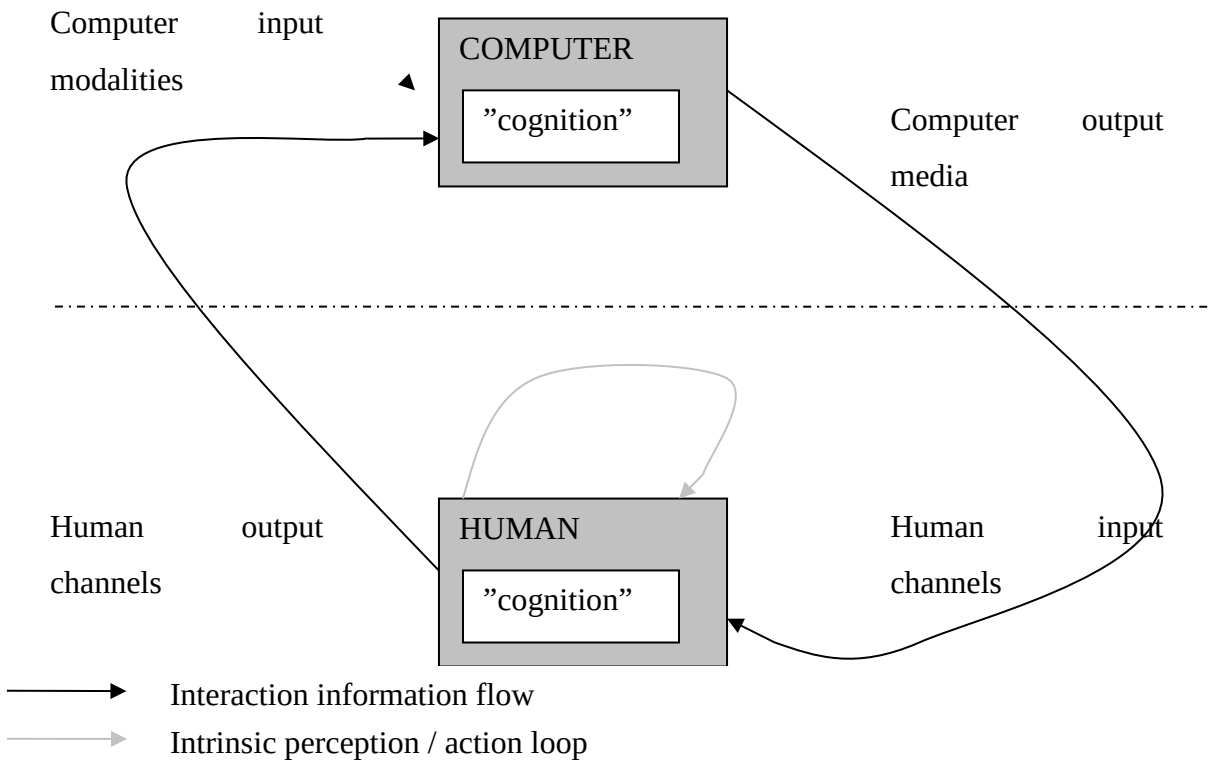
3.3.1 Defining multimodality and multimodal interaction

A verbal definition for multimodality can be defined as (Nigay and Coutaz, 1993):

"Multimodality is the capacity of the system to communicate with a user along different types of communication channels and to extract and convey meaning automatically."

A visual definition of a multimodal system, where both human and system centered views are presented, is depicted below in figure 2. The figure illustrates the concept of multimodal interaction, as presented in the esprit project (MIAMI, 1995) and in Roope Raisamos doctoral dissertation (Roope Raisamo 1999, page 5).

Figure 2. Combined Human centered and system centered multimodal views

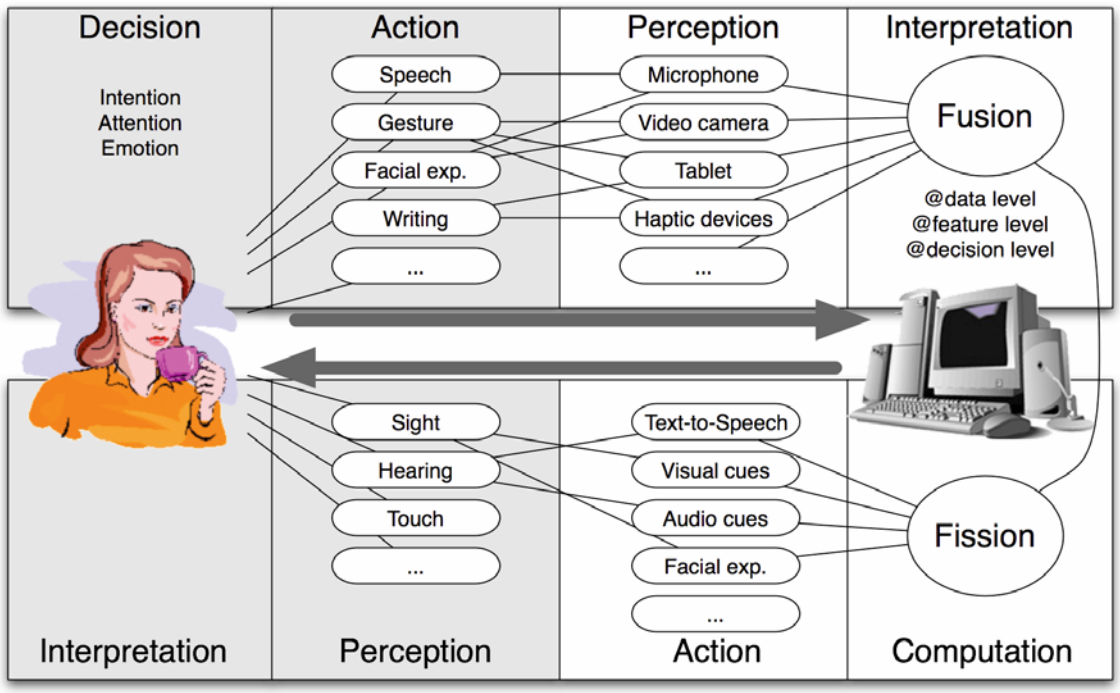


3.3.2 Multimodal principles

As described in the work of Bruno Dumas, Denis Lalanne and Sharon Oviatt (Dumas, B., Lalanne, D. & Oviatt, S. 2009, page 8), the principal dynamics of a multimodal sys-

tem consists of a fusion system, which functions as an aggregation point of various input channels as depicted in figure 3 below.

Figure 3. A representation of multimodal man machine interaction loop



The fission system is the part where a meaningful output is constructed and presented to the user of the multimodal system. The fusion and fission can be seen as a cognitive function similar to what our brains are processing in most of our daily activities, such as when we are driving a car, the brain coordinates the visual data and constructs a mental picture and then commands the legs to push the throttle and keep the hands firmly on the steering wheel.

3.4 The Personal Assistant Model

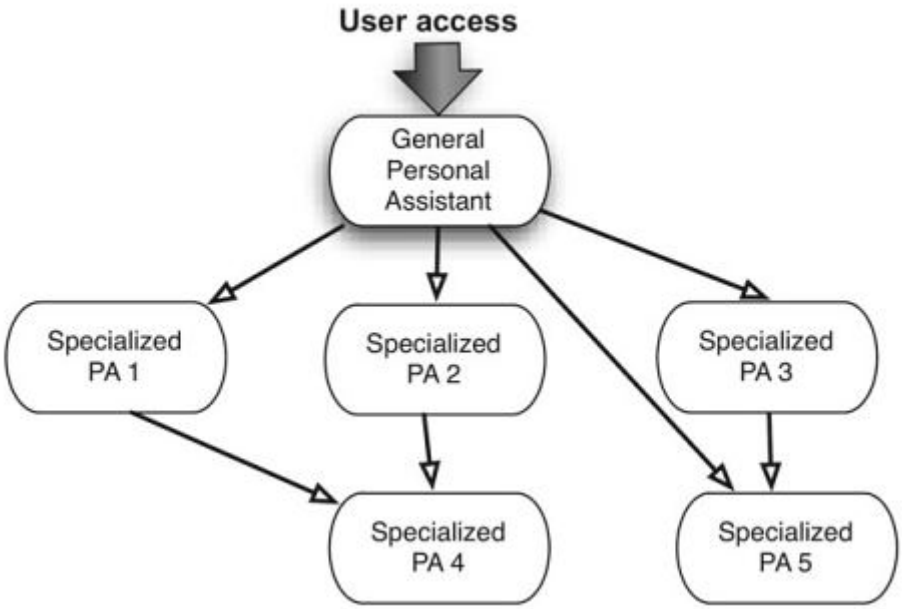
The personal assistant model is defined by William Meisel (Neustein and Markowitz, 2013, p36) as being the following:

“A Personal-Assistant Model (PAM) is software that can take a communication posed in a natural language as speech (as a full sentence or in an abbreviated form), interpret the desired intent of that communication, and provide the user with the desired result as directly and accurately as possible, with one option being a voice response. To the de-

gree that there is context or other input or output modalities available that can help get or deliver the desired result (including entering the inquiry as text and delivering a viewable result), the personal assistant may use those resources."

A conceptual PAM model is depicted below (Neustein and Markowitz, 2013, p37).

Figure 4. A possible relationship between a general personal assistant and a specialized personal assistants (Pas)



3.5 Challenges in applying theories

In order to fully capture a spoken dialogue between party A and Party B, which involves constant feedback through the use of nonverbal and verbal communication between the parties, demonstrates the challenges to entirely record all modalities that might be present in a discourse. This is where the multimodality and multimodal technology can help, namely to break up the discourse into smaller pieces according to the modalities used, with the aim to separate the inputs and outputs used during the discourse and to optimize the inputs for the PA interface (see figure 3).

Due to the multiple uses of senses during a human conversation it will be a challenging task to imitate this behavior technically and provide it as an input to the personal assis-

tant for interpretation. On the other hand, successfully feeding the PA with multimodal inputs can make it smarter and more sophisticated. For instance, if a visual lip reading input and a voice recognition input could be combined, perhaps the overall voice recognition would improve in conditions where there are a lot of background noises.

To summarize, the communication theories explained in the previous chapters highlights the fact that the ordinary human to human dialogue is a complex matter to understand and analyze in depth. It's also important to understand the meaning of modality and multimodality in order to grasp the complexity involved in designing PAs like Siri, Google Now or Cortana, which could be fed with different input modalities or perhaps using the sensor inputs available in the smart phone. Also, the multimodality concept in figure 3 demonstrates that face recognition could be deployed, for instance, if you would have a video call session with your PA, where the camera would input the face movements and the PA could interpret and react on the nonverbal emotions you show with your face movements.

4 METHOD

The Delphi methodology was used as a guideline for gathering of primary data, which is then used to forecast the future of PAs in conjunction with multimodality. The Delphi method wasn't fully applied as such in this thesis. The deviations are mainly that there were no iteration rounds done based on the interview results and the interview data was not shared between the experts. Hence, the interviews were done one to one in relaxed sessions similar to brainstorming sessions, where intuition and creativity are flowing in a free manner. No standard questionnaire was prepared in advance for the interviews as I felt it might lock the conversation or possibly steer the interview in a certain direction. The expert panel consisted of 10 members, which all have subject matter expertise within the area of PAs. All 10 panel experts were interviewed with the aim of collecting the best available primary data, which is as consistent and reliable as possible, in order to secure a solid input database for this study.

5 ANALYSIS AND RESULTS

The primary data was recorded as notes during the interviews and immediately after the interviews. All notes were delivered to each interviewee for inspection and comments in order to eliminate discrepancies in the recorded notes. After this the recorded notes were carefully interpreted and organized in 4 columns and 10 rows. The condensed summary of the expert interviews are presented in table 5 below.

Table 5 The evolution of PA applications

Expert Nr	Current Weaknesses	Future Trends	Modality changes	Future Applications
1	Speak recognition isn't on sufficient level today	PAs will be able to learn new tricks and behaviour on the fly	The preferred user interface in the future will be voice	Voice operated home appliances. (TVs, fridges, microwave ovens etc)
2	Not possible to have a human like dialogue with a PA, because the natural language processing is below 95%	PAs will be the preferred interface to any device or machine	Increased visual dimension. Screens and LCDs project a visual PA image, which can be an animated charachter or human like face. Combinations of voice and projected charachters will emerge.	Smart robots with embedded PAs that can have a human like, interactive dialogue. (Elderly care companions)

Expert Nr	Current Weaknesses	Future Trends	Modality changes	Future Applications
3	Natural language processing isn't at an adequate level yet	PAs get smarter as they are fed with realtime data from intelligent networks connected to Hospitals, Banks, Police.	Moving away from typing and writing inputs on keyboards to PA voice interface.	Specialized and highly personalized PA services within the healthcare and banking sectors. (Elderly people and people with impared vision can get assistance over the voice operated PA instead of typing on keyboard)
4	-	PAs are able to record users data, habits and store it in the cloud. Thus, PAs can make smart decisions and assumptions based on your daily behaviours.	Sensors surround us at home, in the car or in the street with the aim of collecting user data for uploading to the cloud. Kinesics will be used to gather user data	Personalized PA services. (PA knows your commuting routines and is proactively suggesting restaurant or supermarket options that support your grocery shopping list for that day)
5	-	PAs are able to read from the live cam-era feed and interpret and analyze in what mental state the mo-bile phone user is in at any moment based on these inputs.	Increasing the intelligence of the PAs through direct input from camera and microphone. Increased visual dimension. PA will be projected as a hologram.	Implementing face recognition logic to the PA can result in au-thentication and security services (Identification services)

Expert Nr	Current Weaknesses	Future Trends	Modality changes	Future Applications
6	PAs are incomplete, since they aren't capable to offer much of interactivity for the users .	PAs become fully interactive as they will have access to enormous amount of data that is packaged and delivered from high speed intelligent networks.	PAs get equipped with multimodal interfaces(camera) that can enable them to collect more information from the users.	PAs get equipped with new tools that boost their capabilty to deliver personalized services. (Help desk PAs take care of the after office hours. PA would serve in preferred customer language and record the whole conversation in speech and text format. Additionally PAs could ask the customer to show the malfuntioning device to a camera and record the video input, which then can be processed at open office hours.

Expert Nr	Current Weaknesses	Future Trends	Modality changes	Future Applications
7	Speech as a PA interface is always limited due to the slowness of the speech in comparison to visual information showed on a display. The voice interface slowness restricts the future PAs to only deliver simple tasks.	PAs remain on the same level as now and will only be capable to perform simple tasks.	PAs are equipped with vocal emotion recognition capability, which aids them to steer the interactive dialogue with the users.	PAs function as counseling officers to distressed people during crisis management PAs function as ”grown ups” in a helpline service for children Virtual charachters in video games come to life via PAs (Crisis, helpline, gaming)
8	Finance and marketing will pay bigger role in the evolution as there is really no technical obstacles	PAs will be connected to the early phase of the phone call, where they will be connected to tasks.	PAs will increasingly provide more visual information to the users	PAs get connected at early call setup and will handle the incoming call in the best way and interest of all parties (Finance, banking, insurance services, where the voice call will be converted to text transcriptions for legal purposes)

Expert Nr	Current Weaknesses	Future Trends	Modality changes	Future Applications
9	PAs are too learning intensive and it takes too much of effort in learning the key words that enables an effective communication with the PA in order to execute tasks.	PAs are aiding and assisting elderly to prolong their independent living at home	It's not enough to have a nice voice, but PAs must also be adopted to have empathy skillls towards the users	PAs will be integrated in smart home and elderly care environments. (PAs will control vacuum cleaners, ovens and room temperatures over voice interface)
10	-	PAs are designed for offloading simple tasks that doesn't require much logic. This offloading of searching for directions and finding places via voice or natural speech will increase.	Voice modality is increasing in conjunction to PAs, because more tasks are offloaded to PAs. Smart homes will be equipped with microphones and TV screens in all rooms for communication in visual and speech format. Holograms might frequently be used as well.	Help desk services, where PAs can answer questions instead of customers waiting in the queue to get served. (Insurance sector) PA is realized as a physical robot and is programmed to function as a servant for a human. (Smart homes)

6 DISCUSSION AND CONCLUSIONS

6.1 Current weaknesses

Column one is describing on a general level what the main deficiencies and challenges exist within todays PA generation. It clearly indicates that the opinion of the expert panel is that the biggest obstacle is the natural speech processing of the PAs, because it's not on sufficient level. Another interesting viewpoint by one of the experts is that there might be bigger challenges on the marketing and financial side than on the technical side, in order to reach the next generation of PAs. However, the overall picture suggested by the experts is that the Natural Language Processing is an item that is currently the show stopper for evolvement towards the next generation of PAs. The interviewees also point out that the NLP needs to be more resilient for user speech inputs and less error prone. As such, none of the interviewees thought that the NLP issues couldn't be solved in the future. Three of the experts didn't mention any deficiency for the current PAs. Additionally, it should be noticed that no one of the experts mentioned any future dates, when the new generation is expected to be available in the market. The conclusion that can be drawn from the available data is that there are both technical and financial hurdles to be overcome before we can expect a new generation of PAs to become available for us.

6.2 Future trends

Column two describes the expected future trends for the PAs. Here the expert's views are not easily generalized as there are many different trends mentioned rather than a certain common trend. Two experts are expecting that the PAs get smarter, because they will be connected to an intelligent backbone network that is feeding them with information on demand. Another 2 experts are predicting that the PAs will be carrying out the same simple tasks in the future as of today. One expert indicates that the PAs will be able to interpret live streams from cameras and able to work out the mental state of the mobile users based on this input. Another expert thinks along similar lines, while he states that the PAs will be able to make smart decisions based on the user behaviors. Another interesting viewpoint mentioned by one of the experts is that in the future all

machines and devices will preferably be connected via a PA as the main interface. The conclusion to be made here is perhaps that it's as likely that the PAs are not evolving much or if they do evolve they will be connected to a high speed intelligent network as per the PA definition depicted in figure 4. One could conclude that the PAs will be able to make smart decisions based on user input. It might also be the case that this could be realized by the use of a specialized PA that is connected to an intelligent network (see figure 4).

6.3 Modality changes

The third column is reflecting the possible future modality changes. In this column I have specifically analyzed the interview data based on the modalities that were mentioned during the interview. This column is extremely interesting and important for this particular study as it directly captures the future PA changes in regards to the modality usage. Almost all experts agree that the dominating interface will be voice for the future PAs. This is probably not that big of a surprise, since PAs already today, function based on a voice interface. However, many experts also mentioned that there will be an increased emphasizes on the visuals and the experts anticipate that the future PAs will be presenting more visual data to their users. As an example of this, 1 expert was proposing that PAs could use holograms to convey information across to the users. Another interesting view shared by many experts is that the user's emotions need to be captured somehow by the PA. The collection of data from the users by means of sensory equipment also indicated that PAs will evolve towards new ways of collecting user information. Kinesics was also mentioned as a way of collecting user information.

The conclusion that can be drawn is that the expert panel's opinion is that the future PA functions need to be more intelligent. Most of the experts mention that new inputs are essential in order to achieve more PA intelligence, i.e. kinesics, but they seem not to associate the inputs leading to multimodality, where more than 1 input is activated at the same time. Only 1 expert directly connects cameras as a multimodal interface. Nevertheless, another experts opinion is that voice interface will always be too slow, which perhaps can be interpreted that PAs naturally evolve towards multimodality, as only one modality is not capable to effectively serve the user's needs. Perhaps another con-

clusion is that multimodality can be seen as a means for PAs to gain more intelligence. The evidence for this evolution is that the experts repeatedly mention that the PAs will increasingly be equipped with cameras, microphones and preferably even emotional capabilities in order to capture the emotional state of the user at any point in time.

6.4 Future applications

The experts list a variety of possible future applications, but after an initial analysis, there seems to be a trend towards PA applications having a greater presence in a smart home environment. Thus, half of the interviewees share a vision that future PAs will be serving customers in a home like environment. The elderly care was mentioned as a potential ground for PAs to offer and deliver their services basically in some form of a robot. Also, the common home appliances would be PA operated. Another area, where many of the experts shared a common view, was the vision of PAs delivering services in a help desk environment. The help desk services consisted of personalized services connected to the following business sectors: healthcare, banking and insurance. One expert suggested that PA counseling services would emerge within crisis management, where help could be delivered to distressed people during natural catastrophes or war situations. In a way, this could be interpreted as being part of the health sector service. The finance and insurance sector would use the help desk PAs to record the vocal conversation as text transcripts. The transcripts can be used later for legal purposes, in case there are disagreements in what was precisely discussed and agreed, for example, if insurance was bought via PA service. The conclusion seems to be that the experts overall are very aligned in their visions of future PA services. The thinking of the expert panels is that the PA services will be part of the everyday home environment and as help desk services within the banking or finance sector.

7 SUMMARY

7.1 Overview of the thesis work

There was a massive effort done in researching the available background information on the communication theory and the existing multimodality theory. I wanted to describe

all theory in a thorough manner in order to provide the readers with insights to the human sensory modalities, machine sensors, multimodality and PA concepts. Additionally, I hope the material presented in this study is useful for future PA and multimodal studies, as it can be helpful in highlighting the complexities of communications in a multimodal world over a PA interface. Another aim was to produce a straightforward and condensed package of information that takes the reader's knowledge and understanding from basic communication theory towards those of multimodality concepts.

7.2 Evaluation of the Delphi method

The Delphi method delivered valuable input data that could be structured and further divided into categories. Thus, for collection of research material the method worked very well. For predicting the future, the method works as well as the interviewees are able to build a correct view of the future. On the other hand, there is no better information available than ask from the experts. In hindsight, I think this is a weak part of the method as it can only be as good as the expert panel's contribution is. The method tries to compensate this shortcoming via iterative rounds, but I believe that the iterations won't compensate fully in order to make the result data more reliable. Though, I believe that the method produce results that are valid, but if they will be true outcomes in the future is impossible to state for sure.

7.3 Findings and highlights

The findings of this study, after collecting the expert panel inputs and analyzing the collected material, are that the future PAs needs to be more intelligent. The question remains, how greater PA intelligence is achieved. The data collected in this study highlights that the intelligence is likely to be gained by providing the PAs with more inputs, which is a clear signal for PAs future evolvement towards multimodality. Also, the modality changes (see Table 5, column 3) supports that there is a trend towards collecting more information via other means than the traditional keyboard and mouse input. Additionally, the study also identifies the business sectors and the environments that are especially attractive for future intelligent PAs. Another highlight in this study is that the future PAs are probably emerging in smart home or an ordinary home environment.

Yet, another finding in this study is that the banking and finance sector are potential users of the smart PAs.

The communication theories presented in this study also supports that having a discourse using more modalities than mere speech is fuelling the conversations with much more content, for instance, using the nonverbal communication as input during a conversation makes it more vivid and content rich. In other words, having a multimodal face to face conversation is a natural way to use several simultaneous input modalities, for instance, facial movements, body postures, gestures and voice of tone to only mention a few. Providing PAs with the above modality functions will give them much more input data to analyze from their users, which hopefully can deliver the needed intelligence for taking the needed steps towards next generation PA service, i.e. elderly care in smart homes. Also, using machine sensors in order to provide more user information to the PAs can significantly contribute to gaining more PA intelligence. Hence, human eyes function principally in a similar way as a camera. Our ears can be represented by microphones in order for the machines to listen effectively. Perhaps, you can even combine the camera and microphone with other multimodal techniques, such as lip movement reading, in order to achieve even more accurate information to aid the decision making and steering of a user dialogue for a future PA.

As Oksman points out, the mobile is a versatile media instrument in itself, which in practical terms today have many inbuilt machine sensors, camera and microphone. In other words it is loaded with multimodal potentials, but at the moment the PAs only use the microphone as an input and the lcd screen and voice as output. Combining the PA definitions presented earlier with the need of increased PA intelligence, in turn must lead us to envisage a future where mobile PAs are connected to high-speed backbone networks via the special PAs. In this way, the inbuilt mobile PAs can easily be fed with information from various intelligent network nodes, which will enhance PA capabilities and make it smarter. A possible service could be a banking transaction, where the user is identified and authorized by a special PA by user voice recognition instead of inserting a pin code via the mobile touch screen.

7.4 Recommendations for future studies

The foci of this study turned out to be too general and future studies needs to narrow down the scope. For instance, investigating the healthcare PAs in a smart home environment or help desk services for the finance or the insurance sector would be good study objects. There are of course many other directions that can be explored, but the previous mentioned would make good candidates as the experts seems to agree that the future PAs will work in this type of environments.

REFERENCES

Butland, M. 2013, *Achieving Communication Competence: An Introduction to Human Communication,* Kendall/Hunt Publishing Company.

Dumas, B., Lalanne, D. & Oviatt, S. 2009, "Multimodal interfaces: A survey of principles, models and frameworks" in *Human Machine Interaction* Springer , pp. 3-26.

Neustein, A. & Markowitz, J.A. 2013, *Mobile Speech and Advanced Natural Language Solutions,* Springer Science & Business Media.

Nigay, L. & Coutaz, J. 1993, "A design space for multimodal systems: concurrent processing and data fusion", *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*ACM, , pp. 172.

Oksman, V. 2010, *The mobile phone: A medium in itself*, VTT.

Raisamo, R. 1999, *Multimodal Human-Computer Interaction: a constructive and empirical study*, Tampereen yliopisto.

Schomaker, L., Munch, S. & Hartung, K. 1995, *A taxonomy of multimodal interaction in the human information processing system.A Report of ESPRIT Project 8579 MIAMI,* .

Van Servellen, G.M. 2009, *Communication skills for the health care professional: Concepts, practice, and evidence.* Jones & Bartlett Publishers.

Wood, J. 2011, *Communication in our lives,* Cengage Learning.

APPENDICES