Voice and avatar face recognition with focus on familiarity and recall accuracy for use in a contact book designed for illiterates

Alex Patrick Hauge
Stud. Cand. Scient. Medialogy
Specialization: Games
Aalborg University, Denmark
alexphauge@gmail.com

Christian Bødtkjer Jørgensen
Stud. Cand. Scient. Medialogy
Specialization: Sound and Music
Aalborg University, Denmark
cjoer86@gmail.com

ABSTRACT

In the world there is around 799 million illiterates [1], when illiterates use a digital interface they are normally restricted by the highly text based interfaces typically used. This limits their ability to navigate and use the functionality seamlessly without reading aids, instead they use a combination of tactics and assistance from literate people. This paper focuses on how voice and avatar faces can improve how an illiterate uses a contact book. The results shows an increase in recall accuracy due to an increase in familiarity within both avatar customization and voice recognition. For testing an interface, a prototype was developed for an Android mobile phone. This contact book application included a way to sort through all contacts by a visual search feature. Within the application avatar customization and voice recognition is the main focus of the concept.

Author Keywords
Illiterates; Voice Recognition; Avatar Recognition; Mobile User Interface;

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: User Interfaces

INTRODUCTION

It is estimated that around 799 million people are illiterate [1], each facing a problem of navigating their daily lives when faced with text in different forms. Simple tasks as interacting with an ATM or navigating a phone are tasks that literate people perform seamlessly, this is also the case for illiterates that have learned tactics to cope with the text-based interfaces. These tasks can prove hard or almost impossible to perform for illiterate people, as they are not able to read the information needed to navigate the systems. As the mobile phone is integrated more and more in society, an increase in illiterates using mobile phones will occur. To cope with the problems that they encounter, illiterates have found different tactics to deal with the fact that they cannot read. A tactic applied by the illiterates in their paper address book, is using spatial location, number, color, page number and doodles in order to help them identify the contacts in the address book [2].

The problem is that most interfaces are designed for literates, therefore making it hard for illiterates to find a usable interface that take their disability into account. In a normal contact book on a phone, the contacts are displayed with a name and a number alone. The focus of this paper is to investigate how to create an illiterate user friendly contact book on a phone, in order to find a way for illiterates to ease the use of the contact book. We are in this paper leveraging the coping mechanism illiterate's use on paper, by introducing spatial search on the contacts and allowing disambiguation through speech in a prototype. For this we assume that there is no difference among literates and illiterates with regard to facial and voice recognition. In this paper we define a contact book as an augmented people manager, that allows one to organize and store information about ones contacts. There might be a conflict with recording voice samples from a conversation because of laws, an alternative that still gets voice snippets could be to record using the phone in a meeting with the contact. Several voices could appear in such a voice snippet when recording in environments with other people.

The paper contains sections about background knowledge, test methodology and results and lastly a discussion.

BACKGROUND

Illiterates use of Mobile Phones Memory

There are different levels of illiteracy, functionally illiterate, illiterate, semi-literate and low-literate. Functional illiteracy refers to people who have problems writing even though they attended education [3]. Illiterate people have basic numeral skills, they can to some degree read and write numbers, but they cannot read and write. Semi-literates is the class of users who can understand numerals, but have difficulty reading and writing [4, 5]. Low-literate people generally forgets the alphabetically organization or do not know the alphabetical order [6]. In this paper we focus on the illiterate level. We use a definition of illiterates as people that are not able to read and write words, but able to read and write numbers by Friscira et al. [5].

Illiterate are forced to find tactics that can help them cope with the fact that they cannot read or write, if the illiterate per-

CHI'13, April 27–May 2, 2013, Paris, France.

son cannot obtain help from a literate. They use cope strategies such as memorizing the order of contacts in the contact book or searching by the first letter.

Illiterates are faced with different problems when interacting with digital text based interfaces, but as found by Lalji and Good [2], they still desire mobile phones to either keep in contact with family and friends or for society standards as described by Knoche and Huang [7]. As described in Medhi et al. [8] there are over 4 billion phone users in 2008, 40% of these are from non-developed countries, in the least developed countries 41% of the population is illiterate.

The illiterates disability to read often force them to live in a more predictable way, simply doing the same thing instead of asking what something means, claims Chipchase [1]. [8] describes how illiteracy is viewed differently. In developed countries shame is attached to illiteracy whereas in undeveloped countries it is not. Despite this, the article [2] describes illiterate people from India that do not feel clever enough for interacting with a mobile phone interface designed for illiterates, in a user centered study.

Illiterates use a variety of different tactics when dealing with the process of saving and retrieving phone numbers. Illiterates memorize the most important phone numbers [2], whereas others are written down in a notepad or diary as described by Joshi et al. [6], and identified by handwriting, page number, marks, color or position on the page [1, 9]. They tell the numbers apart from spatial arrangements location in the notepad or from use of colors, patterns and doodles [7]. Some illiterates and semi-literates recognize a phone number or name by turning it into a symbol or image as described by [2, 6, 7, 9]. For illiterates adding a new contact on a mobile phone, they most often depend on a literate person as described by [6] and Ahmed et al. [9]. These helpers assists the illiterates with entering phone numbers, but also assist in the process of learning and memorizing different functions [1, 4, 7], and updating information in the contact book [4].

[6] found that low literate users saved the name of the contact with a number in front, to ensure the location in the contact book for later use, similar [7] found illiterates using numbers instead of names for contacts. The tactic of saving a contact with a number in front for the purpose of order, is similar to the tactic by illiterates when searching for a contact in the contact book or recently called, by remembering the position of the contact on the phone [6, 9] or using the call log as a contact book [1,7]. Contrary to using the order of the contacts in the contact book or call log, [6] found low literate users saving contacts in relation to the location of the contact, such as town or country.

Another way some illiterates deal with phone numbers is using SMS as storage with a phone number and a name. The information will be read aloud by a helper [5]. When the illiterates receive a message, instead of writing, they call the person. Unlike using a contact book, call log or SMS, some illiterates do not use the mobile phones features, but instead calling from scratch each time as found by [8] in India.

A tactic applied is remembering the two to three last digits of a phone number when searching the contact book [7,9], some use the frequency of letters in the name when searching [9] or patterns in the phone number [6]. Another tactic is identifying a contact by the first letter and then use other features for differing in case of several contacts with the same first letter [7].

When navigating the mobile phone, it was found for low literate people that they coped best with a linear structured interface, followed by a cross linked navigation structure as described by Chaudry et al. [10]. Similar it was found by [8], that a hierarchical navigation structure proved hard for illiterate people. Function that was buried to deep was found to be less discoverable [8].

Even though illiterates are faced with this barrier between them and a mobile phone, they find ways to cope with it. Illiterates are not frightened by mobile phones and find the same need for them as literates do. As described in [7] they want a mobile phone because everyone else got one and are generally desired by illiterates [6]. Both illiterate and literate have the need to keep in contact with family and friends, something that the mobile phone help provide. At the moment the illiterate's uses a combination of tactics and pen and paper to cope with the text based contact book on mobile phones, that does not support their situation.

The illiterates memorizes the most important phone numbers, this still leaves a lot of contacts that are based on text which they cannot read. It is necessary to find a way to design an interface that augments text use and that is able to convey the needed information to the illiterates.

Visual Identification from Avatars or Faces

When recognizing a person you have seen before in a picture, that includes the face, the name might still be unclear. You will still be sure that you have seen the person before, which can lead to other identification methods than facial recognition, such as clothes, location, time or scenery the picture includes. The face contains a lot of social meaning and information about emotions, which the brain is highly optimized to detect and then identify certain traits. Or and Wilson [11] shows that the brain detects a face and then tries to figure out who this face belongs to through identification.

Facial recognition is the act of recalling that you have seen the person before. The act of identification lets you remember the name or context in which you came to know the person. Sun et al. [12] supports that the brain detects structural setup of available facial features and then goes on to social meaningful features, such as gender.

When exposed to a certain range of facial features, the brain becomes adapted to that specific range, which Young et al. [13] touches upon by examining the Cross-Racial effect. The Cross-Racial effect is what cause asian people to better tell asian people apart because of prolonged exposure, while caucasian people are much worse at telling asian people apart if they have not had exposure. Due to this exposure to family and friends, which should be people you see often, should become easier to recognize and tell apart [14].

Elftmann [15] writes that face recognition is an easier recall task than remembering a password. In [15] a study showed schoolmates being able to recognize each other after a 35 year period, with over 90% accuracy. Even unfamiliar faces, which people trained to remember in a two week period, gave a high success rate over time having the longest period of 5.4 months after which the total accuracy was 72% success. This tells of how important familiarity is, with the persons involved, as emotions can help improve the recognition and identification [16]. Though everyone has experienced someone getting a new haircut, making it hard to recognize them, which goes to show that hair style and color has a major impact on recognition [17].

Belle et al. [18] and Wang et al. [19] describes two different approaches to face recognition. The dominant one, holistic percept, is using all features present on the face to generate a holistic whole. The other approach is feature-by-feature recognition, which is a lot harder and more prone to errors as it checks each feature by itself. Gavrilova and Yampolskiy [20] found that facial recognition and identification methods described can apply to avatars. They used algorithms for face detection to identify avatars in various scenarios or cases, such as in a game, robots in real life, etc.

Avatars are used for depiction of identity between players in the popular Xbox and Wii environments. No sorted contact manager exists for these environments, which does not enable using the avatars as a retrieval aid. In a contact book, avatars can be used as a retrieval aid to help recognition.

Using avatars can be helpful in cases where no access to pictures of the individuals, Berg et al. [21] also used avatars in a contact manager. They sorted contacts through prioritizing, which the owner of the phone communicated with more or less, by having the depth of each contact be either closer or further away.

Having too many features to customization of an avatar might not be for the better, as it complicates searching through the features. It also makes the creation of an avatar more time consuming as you would have to go through several options to get to the right one. There is a fine balance between adding too many features and getting just enough features for a good enough representation of a person and fast search results.

The Ability to Identify a Person by Voice

The process of identifying a person from their voice is not only a matter of recognizing the audio, but a process can use both face and voice input by Campanella and Belin [22], or seperately as written by Joassin et al. [23]. The reason for this is the mechanics behind speech as described by Lander et al. [24], that connects the facial movements of speech together with the voice, which the human being takes advantage of when identifying using voice.

When identifying a voice it is divided into two sections as described by Kuhl [25] besides the face and voice. The voice identification cares for the "who" and the speech perception the "what", these two are integrated.

It is also possible to recognize a voice without facial input, such as over the phone, in those cases people try to discern gender, age, emotion and familiarity [22, 25], but voice recognition is easier when presented with an associated face [22,23]. When presented with a non-associated face when doing voice recognition, the accuracy of voice recognition is lowered [22, 23].

This shows that it is not possible to ignore a face when presented with it [23]. When identifying a voice it is possible to do so for a silent moving face (the facial articulations of speech with no sound), the things that are important for identification is not the content of the sentence but the non-verbal. Such features as how the word is stressed, emotions, attitude and manner help identify the speaker [24].

Trying to identify a voice after a period of time, will cause a lowering in accuracy of identification, Clifford [26]. For unfamiliar targets it was lowered to 50% after 1 week and 43% after 2 weeks. The participants were trained with one live voice and when tested they had to identify from a series of 10 voices. McGehee [27] found that recognition of unfamiliar voices declined to 80.8% after one week and 68.5% after two weeks, a higher accuracy than described by [26].

The participants training was a 56 words passage, which had been read aloud for them. In the test they had to identify from 5 different voices which included the 1 trained target. The accuracy of voice identification is not only affected by the duration between stimuli, but also from the age of the subject [26], subjects with the best voice recognition were people in the age of 20-40 years. This might be due to general memory [27], that also plays a role in the recognition of voices after a period of time.

The effect of familiarity on stream segregation (The cocktail party effect, in which you can still listen to one voice when situated in an environment with multiple voices) researchers Newman and Evers [28], found familiarity had no significant effect when stream segregating. [28] found there are different levels of familiarity, such as being familiar with the voice or being personally related to the individual behind the voice. Being familiar with the person should improve voice recognition [28].

The effect of familiarity on voice identification was investigated by Yarmey et al. [29], the people were asked to provide several voices and faces of people, divided into three levels of familiarity, high, moderate, low and unfamiliar which they provided the people for. In the article by [29] they found that there was a significant difference in accuracy between high and moderate, and low and unfamiliar, where the low and unfamiliar produced significantly more errors in identification than the group with high and moderate familiarity.

The GSM (Global system for mobiles) has different standards for sampling and compressing signals [30]. For a transmission of a signal without compression, it requires extra bandwidth causing low bit speech coders to be the standard in international and private communication systems [31]. There are speech coders covering 16 kb/s to 8 kb/s and even as low as 2.4 kb/s, causing the loss of higher speech complexity [31].

The latest trends is coding in the range of 6kb/s to 2kb/s but using speech specific coders for extraction of speech specific information to compensate for the low kb/s [31].

From the research we chose to design a linear structured interface for our contact book, that includes numbers and avatars to further help the illiterates disambiguate between contacts. A spatial search is designed to make searching more accessible to illiterates. To further disambiguate between contacts a voice recall function is designed.

PROTOTYPE

We developed an Android prototype contact book for smartphones, to illustrate the usage of the different features. For the design it is known that a contact book will contain around 80 contacts [6] and be used in an everyday scenario. Illiterate users will recognize frequently used contacts by their phone numbers, but a lot of entries will not benefit from this and not contain text. The main grid is meant to be ordered by last called, which some illiterates uses to disambiguate who they talked with last.

This prototype allows for creating visual avatars to represent the specific contact, see Figure 1.

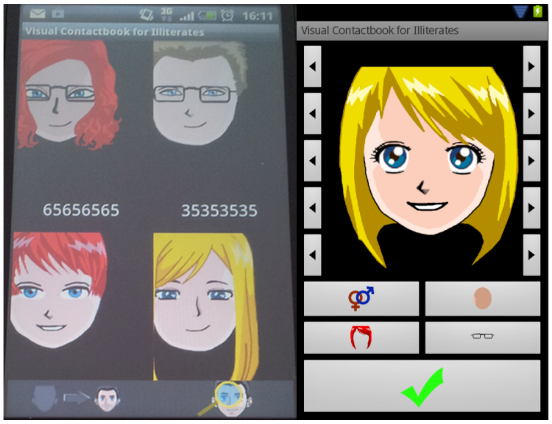
16:11
Visual Contactbook for Illiterates
65656565
35353535
Visual Contactbook for Illiterates

Figure 1: To the left the main screen grid of the prototype, to the right the customize avatar screen of the prototype.

The image of the created avatar is saved with a name and phone number of the contact. The search function is based on facial features. In the search screen a grid of contacts is displayed, as different features are chosen, the grid is updated with the contacts that meet the feature requirements. When pressing a contact you enter the profile of the contact. The contact profile screen displays an image of the contact avatar representation, a button for voice playback.

Both tactics can be used in symbiosis with our way of helping them remember the person through avatar representation and voice identification.

TEST METHODOLOGY AND RESULTS

We carried out a total of six tests, three to investigate sound and two for the avatar representation of the contacts. The last test is one concerned with the usability of the interface in the contact book.

Voice-Avatar Identification Test

This test is made to find out how well voice can be used to aid illiterates improve identification in a contact book with visual avatars, for unfamiliar voices in relation to an avatar face. We chose unfamiliar voices because it would resemble a worst-case scenario in recognition.

For this, two groups were formed, one concerned with voice-to-face and another with face-to-voice, inspired by the test method from [23]. The two groups were formed to test the voice recognition impact regarding changes in situaion.

This between subject setup means that the group with voice-to-face are first presented with a number of voices and then asked to identify the voice related to the one face that they are shown, the reverse applies for the face-to-voice group. Before testing, the participants in each group have to be trained on the avatars and related voices, this is done for both groups.

We recorded voices for the test, in Aalborg using a Samsung Galaxy S2. Each person which voice was recorded was told to say "Hvordan har din dag været?" in Danish, which translates to "How has your day been?". The reason for keeping the sentence the same, is to focus on the ability to recognize voices, even though they say the same sentence it opens the possibility for small cues to appear. The average length of the voice snippets for the male targets is 2.33 seconds and for female targets 2.58 seconds.

We created avatars on the homepage called FaceYourManga.com [32], the reason for this is that the style used on this page is similar to the one intended for the project, a style between human realistic and cartoony.

When the test starts for either the voice-to-face (VF) participants or the face-to-voice (FV), two facilitators are present. The test is divided into female and male avatars to keep gender from being a factor in recognition. Before testing on female and male avatars, both groups VF and FV are trained on 6 male avatars with voices for male targets and 6 female avatars with voices for female targets. All of both female and male voices are unfamiliar to the test participants. The training consists of them watching the face and hearing the related voice, two times for each of the 12 targets. The test presented the participant with either four voice to one face in VF (one set) or four faces to one voice in FV. Each participant is tested on three sets for both female and male avatars after the proper training. In both groups when presented with the four same gender voices or faces, there is one target together with one previously seen target and two distractors which the participants had never seen or heard. The voices and faces in the sets are randomized in order to try and rule out any pattern, that could help the participant guess the next one. For this test 16 people participated, 3 women and 13 men, all university students. Each group, FV and VF, has 8 participants.

Results

This test is performed with unfamiliar targets. The test data is divided into two groups FV and VF, within each of these, the data is divided into a correct answer, wrong target and wrong distractor. The wrong target is when the person chooses wrong, but chooses a previously either heard or seen target.

The wrong distractor is when a wrong choice is made, by choosing a distractor. Figure 2 and 3 shows both FV and VF

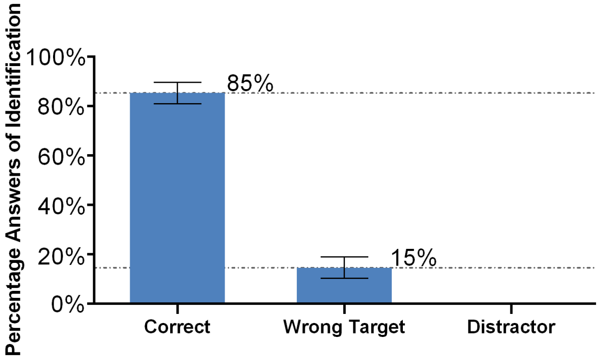

Figure 2: Correctly identified avatar profile voice and errors, from a set of 4 faces, the face of the avatar profile, 1 previously seen and two distractor faces never seen before.

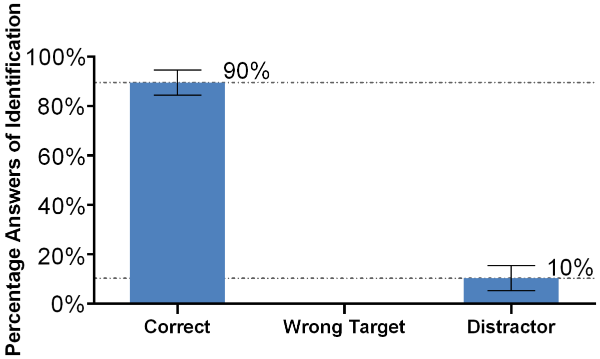

Figure 3: Correctly identified avatar profile voice and errors by having the avatar profile image shown. Identification is done from a set of 4 voices, the voice of the avatar profile, 1 previously heard and two distractor voices never heard before.

have a low error rate below or equal to 15%, showing that it is possible to identify the avatars with sound. When the participants is presented with the four voices in voice-to-avatar face they show that it is harder to recall all trained voices.

Even though the FV group chooses previously seen targets when performing an error, the VF performs with a lower error rate then FV with a difference of 5%. From this test it can be said that it is feasible to use avatar and voices for recognition features within the concept of the contact book.

Sound Quality Test

This test is concerned with the impact that sound quality has on voice recognition. The main reason for this test is that we assume recording of voice for the application, over the phone from a conversation. The reason for this is that sound quality may vary from phone to phone. This is relevant due to the growing use of mobile phones, because by reducing the quality, the bandwidth and power consumption can be lowered.

The stimuli used in this test is the same audio that was recorded for study 1, being unfamiliar voices. As this test is focused on the impact of sound quality on voice recognition, the audio has been resampled. We recorded voices originally at 44khz, 16 bit PCM Mono. The three qualities used for the test is 8, 24 and 32 Kb/s. These qualities are chosen due to the low-bit speech coders in GSM [31], covering from ADPCM (high 32 kb/s) to medium and low at 16-8 kb/s.

The test is divided into a section with female voices and one with male voices given a number for identification. Each participant is tested on their ability to recognize voices resampled at three different qualities using unfamiliar voices.

The reason for not using a name, is that people might have relations to that name, increasing their recognition. The test is setup as a within subject test, meaning that they will be tested on each quality at each location. Before the test starts they are trained on the voices with the specific gender they will try to identify and after the test, training follows on the opposite gender. The training consist of 6 male or female voices, that they hear twice. In each part of the test, dependent on gender, they are presented with 9 sets of female or male targets, in each set there is a target voice, a distractor and a previously heard voice that is one of the 6 voices heard during the training. For this test 10 people participated, 10 men, 20 to 30 of age, all university students. The three qualities that are used for the test are 8 Kb/s, 24 Kb/s, 32 Kb/s and 44 Kb/s. The 44 Kb/s is used for training.

Results

The success rate for each quality is displayed below.

- High (H) = 51 correct (85%)
- Medium (M) = 46 correct (76.7%)
- Low (L) = 42 correct (70%)

It is not possible to locate a point where it is no longer possible to recognize with a higher then above average, a graph is made using the information from the qualities and assuming that 0 Kb/s is equal to 0% recognition. The data from the different qualities is displayed in Figure 4. The graph in Figure 4 gives an indication of where the cut-off point is for quality in relation to voice recognition. The tendency for recognition goes down correlating with quality. As we can see in the graph the recognition is still high above chance with a quality as low as 8 Kb/s. This indicates that the voice recognition feature is usable even in cases of poor quality sound due to phones or infrastructure.

Study of familiarity effect on voice recognition over time

The goal of this is to investigate the impact of familiar and unfamiliar voices when trying to identify them after a time period.

The stimuli used in this test divided into two sections. For the unfamiliar targets, 4 voices from study 1 is used, two male and two female voices. For familiar targets the participant is told to call 4 friends or family, letting them say the sentence "Hvordan har din dag været?" over the speaker, while it is recorded using the Samsung Galaxy S2. The average length of the voice snippets for the familiar targets is 2.27 seconds.

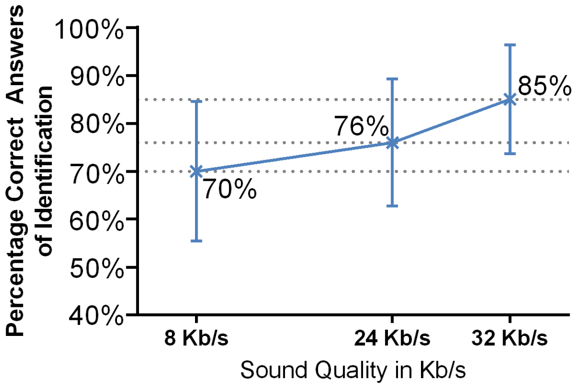

Figure 4: The participants recall accuracy when trying to identify unfamiliar voices of different sound qualities. The x-axis is measured in Kb/s and y-axis in percentage of correct identification answers.

The test is divided into two sections, training and recall, as the test is performed over time. The first step is where the test participants are familiarized with the targets, listening to all the voices 5 times as part of training. This is also the case for the group with familiar targets. For the unfamiliar they are trained on two female and two male voices, for the familiar it was not possible in some cases to obtain two voice samples of each gender.

The next section of the test is two days later, where they are called back, to be tested on their ability to identify the targets. The test is a between subject test, each participant is run through 14 sets, two of them with no targets in. Each set contains three voices, containing a target and two distractors. Each target is counterbalanced so that it is presented in all positions, similar to study 2 with sound quality. This applies for both familiar and unfamiliar. The unfamiliar targets are given a number for identification, because of the same reasons as explained in study 1, and the familiar the name of the target. The test ends with a questionnaire which the participants fill out to help specify extra details for the test.

- Have you talked with any of the subjects from the first part of the test since it was recorded?
- If you have, how often and how long?
- What helped you recognize the voices?
- What proved hard to recognize and why?

For the familiar group it was necessary to find out if they had talked with any of the targets during the two day period, as this will affect the results. For this test 18 people participated, 18 men, all university students. Each group, familiar and unfamiliar, has 9 participants.

Results

The hypothesis that is answered in this test is the following.

- **Hypothesis** - When matching identity from audio input, there is a difference over time between participants that are familiar with the targets and those who are not.

We performed a two tailed t-test on the two samples, the familiar group and unfamiliar. Some of the participants answers, in the familiar group, have been removed due to the fact that they had talked with some of the targets during the 2 day period, therefore making it necessary to rule those out in order to keep the time period equal. In Figure 5 we show the amount of correct answers per participant group.

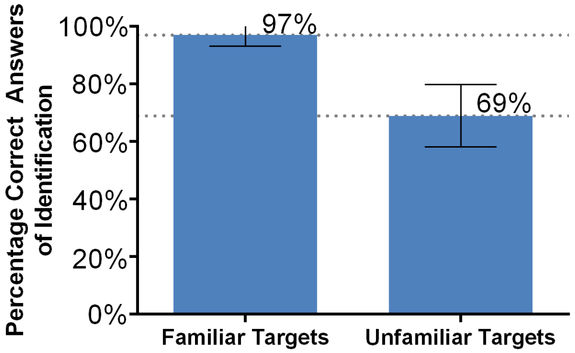

Figure 5: Amount of correct identification answers from each participant group, when trying to identify either a familiar or unfamiliar target voice after a period of 2 days.

The participant group who had familiar voices performs significantly better ($P < 0.01$) in recognition of voices over time as expected. As the contact book is designed with familiar contacts in mind, it proves that voice recognition is a feasible feature for identifying a contact. Even in the case of unfamiliar targets it allows for 69.04% success rate in the same time period.

The different features that they answered as important in the qualitative questionnaire are arranged in a diagram in Figure 6. The three most important features according to the test participants are pitch, pronunciation and background noise cues. This gives a good view of what elements they focus on

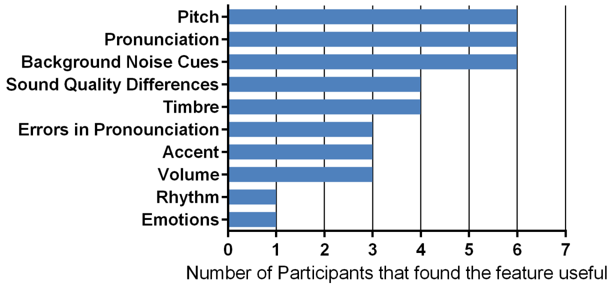

Figure 6: Each number on the x-axis shows how many participants mentioned the feature in the question “What helped you recognize the voices?” as an important feature for recognizing.

when recognizing voices and what to focus on when recording the voices. The fact that background noise cues play such an important role, indicates that to use voice snippets for disambiguation in a contact book, it is necessary to allow background noise to enter the microphone instead of filtering it out.

Investigation of whether facial recognition methods apply for avatar face recognition

The focus of this test is to see if facial recognition methods, feature specific identification or holistic identification, is used for avatar face recognition. We need to conduct this test to know which facial features needs more attention than other features when designing our avatar costumization. In this test, participants see a face that is masked at a certain location to either enable feature specific identification or holistic identification. To enable feature specific identification, we needed to make sure at least one feature was masked to break the flow of the holistic percept, so the participant would need to look at each feature individually to recall. The holistic percept on the other hand, needs to have all inner facial features visible to generate a whole complete percept.

The test participants is divided into two groups, a group with feature specific identification and one with holistic identification. The test setup is similar to [18], but in this test the participant starts by seeing a series of 6 avatar faces that the participants need to remember. This was shown 3 times with a 2 second period for each avatar. Each participant saw 12 sets, each set contains 6 avatars that are masked according to the participants group. Of the 6 avatar faces, there are 5 distractors and 1 target. When the participant is presented with the 6 avatars in each set, they are asked to find a specific avatar from the previously seen avatars, making it a recall task where they remember each avatar by number. The reason for using numbers and not names is explained in study 1. Within each group, feature identification and holistic identification, the groups are divided into two, one with female avatars and one with male avatars, having 4 participants per gender. For this test 16 people participated, 1 women and 15 men, all university students. Each group, feature-specific and holistic, has 8 participants.

Results

The hypothesis that is answered in this test is the following.

- **Hypothesis** - There is a difference between holistic and feature-specific recognition of avatars.

The hypothesis is rejected, though not entirely convinced, through a students t-test (P = 0.053). The high correct recall still shows that feature specific and holistic identification works for an avatar. We did an analysis on the control variables looking at distinctiveness of each avatar in each set, which showed that more distinctive (66% correct) avatars gave a higher recall accuracy compared to lower distinctive (53% correct) avatars.

This is demonstrated through the summarization of answers in Figure 7. In [18] the results for accuracy for normal observers with human faces found that there was a significantly difference between holistic and feature specific. The holistic have a significantly higher recall with approximate 97% than feature specific with approximate 89%. The result that holistic scored higher than feature specific is similar to our test in which holistic scored higher than feature specific, although not significantly.

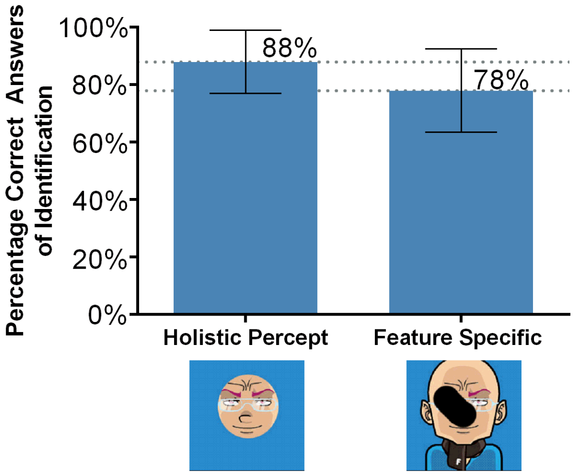

Figure 7: Percent correct answers and SEM when trying to identify avatar faces in either holistic percept or feature specific identification each test set consist of 1 target avatar face, 2 previously seen and 3 distractors never seen before.

Study of familiarity effect on avatar recognition over time

The test is designed in order to see if there is any difference in the ability to remember an avatar over a time period of two days, with focus on the impact of familiarity.

For this between subject test there are three groups, a control group, a group with unfamiliar avatars and a group with familiar avatars. The control group is presented with premade avatars, whereas the unfamiliar group will create their own avatars from a series of pictures of people. The familiar group creates avatars from pictures of their friends or family. In each group the avatars that are created are divided into two males and two female avatars, a total of four avatars, keeping the number of genders the same across the groups. Each avatar in each of the three groups is created with the same possibilities in features.

The test was set up to be carried out over two days, the first day they are trained on the avatars, created or premade, dependent on the group that the test participant is in. They see each avatar 3 times for 5 seconds. Each avatar is given a number as reference, as described in study 1, except in the case of the familiar group. After two days, we tested their ability to recall the avatars. The test participants are shown 16 sets of avatars from the correct group, each set contains at least two distractors and up to two targets, totalling 4 avatars. The test participant is asked to look at the avatars, and see if he/her can locate any of the four avatars that they were trained on, in the set. We ask the participant to answer "what made it easier to recognize the avatars?" to gain further information of how they deal with recognizing the avatars, such as specific features or tactics. This question was answered very freely with us only noting keywords such as hair style, hair color, eyes, etc. words which could help us locate important features. For this test 24 people participated, 24 men, all university students. Each group, control, unfamiliar and familiar, has 8 participants.

Results

Each group has a significant increase in percent correct over the previous group in the order, control, unfamiliar and highest scoring familiar group (Control/Unfamiliar P < 0.05, Control/Familiar P < 0.01, Unfamiliar/Familiar P < 0.01). The means and standard error for means for each group can be seen in Figure 8. This could be coupled with social meaning given to the avatars, by having the participants create the avatars from people they interact with on an almost daily basis. The reason for a lower recall accuracy in the control group, might be due to the lack of social meaning given to the avatars through creating them. When asked about which features the participants used to disambiguate, 20 out of 24 of them answered hair style and color where as other features scored around 5-6 people mentioning them as seen in Figure 9. This clearly indicates that in this scenario with limited features the hair style and color played a significant part of helping the participants in recognizing the avatars just as in [17]. Though some participants noted they used the expression associated with the avatar they made, such as looking happy or sad, to associate the avatar to the person they were mimicking.

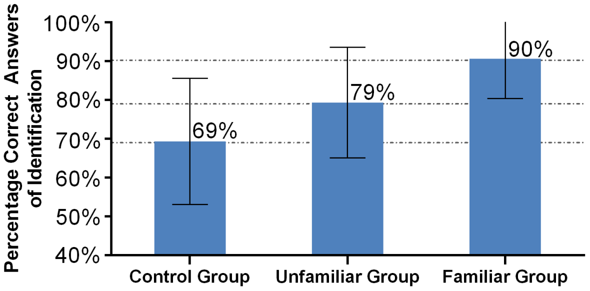

Figure 8: Correct identified avatars with mean in percent and SEM, with different levels of familiarity. The participants create 4 avatars using the application for recognition in the test, this applies for all groups except control group who had premade avatars.

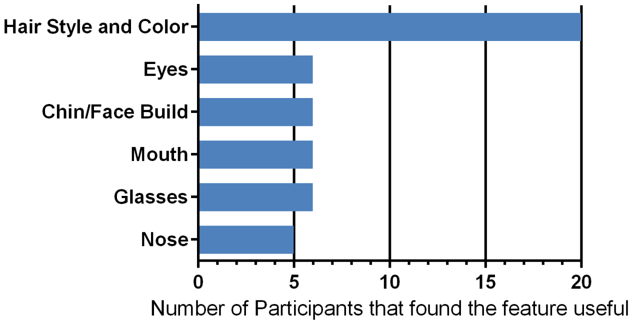

Figure 9: A diagram of the features that was found most important. Each number indicates how many answered that feature as important for recognizing avatars after a period of 2 days.

User Interface Test

To test the usability of the user interface, we tried to simulate the situation that illiterates are in, by using Chinese letters, thereby preventing the participants from being able to read text. The test participants were not illiterate as it was not possible to gain access to illiterate people. This means that it will not be possible to get the effect from the tactics that they use when interacting with digital interfaces.

The test is divided into two sections, in the first the participants carry out a number of tasks on the prototype contact book in a think aloud fashion, in the second, a situated interview. The tasks are designed to test the different interfaces and functions of the prototype, during the task the participant is asked to think aloud when performing the tasks.

- Create a new contact (Female, blond hair, long hair, green eyes, glasses) with the phone number 38651942.
- Locate the contact in the contact book.
- Use the search function to locate a contact.
- Play the voice of the contact.
- Call the contact and then delete the contact.

The test was performed as an iterative process, where flaws in the design were changed after each participant, allowing for more errors to be discovered. The contact book was filled with several premade contacts, trying to provide an equal amount of females and males, to simulate a more realistic use of the contact book. After performing the tasks on the phone with the prototype, the test participant carried through a situated interview.

- Age?
- What did you find difficult and why?
- What could be improved and how?
- What would you use the app for, in what situation or case?

For this test 3 people participated, 3 men, all university students.

Results

During this qualitative usability test a total of three literate participants were tested.

The three subjects performed with few errors when performing the task, errors that would properly not occur after a couple of tries. They were able to use the prototype and its features with relative ease, capable of creating, searching and calling a contact. As in any interface there is a learning curve.

For a more relevant usability test, it is necessary to have a test with illiterates, in order to understand their needs and the effect of the strategies they apply.

DISCUSSION

In study 1 we tested how well an accompanying voice to an avatar face would function when identifying an unfamiliar voice. The test supports the theory about it being possible to use a voice to help identification of an avatar in a contact book. An interesting occurrence is the fact that the VF group when performing an error chooses a distractor, whereas the FV group chooses a previously seen avatar when performing an error. This could indicate that the test participants are better at recalling the visual information, so when in doubt they choose something they have seen before. With the VF group this might indicate that the voice recall is not as good as visual recall, since the distractors are chosen when performing an error.

Study 2 was a test of how much an impact sound quality had on the voice recognition. The theory was that the lower the quality, the higher the error-rate. The test supports that the voice recognition success rate declines as sound quality is lowered. For a more accurate representation it is necessary to test with more intervals in quality below 8 Kb/s, from the current data it is not possible to estimate the cutoff point.

In study 3, the effect of familiarity was used as a variable over a two day time period to see if it had an impact on recognition accuracy. The study showed that the group of participants with familiar voices performed significantly better ($P < 0.01$) in recognizing voices over time as expected. The subjects was also asked questions which helped understand the tactics that the subjects used when recognizing voices. The greatest focus for voice recognition was in pitch, pronunciation and background noise cues.

We tested to see if facial recognition, feature specific and holistic identification, would work for an avatar in study 4. We found no significant difference between each approach. Similar to the test results in [18] which tested with human faces, the feature specific had a larger drop in accuracy than the holistic approach with avatar faces. Though [18] had a significant difference we can argue why we did not. The reason could be the difference in level of detail in a human face and an avatar face, as increased detail in the human face leads to a higher holistic performance than the lower detail in the avatar face. This could make feature specific identification approach holistic identification accuracy for avatars. Because the P value is equal to 0.053 it might have an impact to have more participants.

Study 5 was a test that used time as a factor to see how much influence familiarity would have on recognition accuracy. The results showed that the control group which only had premade avatars, had a significantly lower accuracy than the unfamiliar and familiar group. The difference between the control group and the other two, was mainly the introduction of the participants creating the avatars themselves, from either an image of an unfamiliar subject or a familiar subject.

Because we did not test the identification accuracy right after training for study 3 and 5, we can not see the difference between starting accuracy and the accuracy two days later. But we assume a significantly higher recall accuracy than what they achieved two days later. Study 1 and 4 tested recall accuracy immediately after training with a larger training sample that might worsen performance. Because the training is done on unfamiliar targets in study 1(Avatar face-to-voice 85% and Voice-to-Avatar face 90%), it can help indicate a start recall accuracy for the unfamiliar groups in study 3(Unfamiliar group 69%) and 5(Unfamiliar group 79%) after two days.

This introduced play as a variable which strengthened the accuracy through each group by approximately 10%, meaning the familiar group scored highest with a correct accuracy of 90% while the unfamiliar had 79% correct. A t-test showed that the two groups scored significantly different, supporting the hypothesis that having familiar targets improves accuracy in recognition. When looking at only the unfamiliar group and the control group it can be seen that there is a significant difference, supporting that play has an important role in improving the accuracy of the recognition, because both groups use unfamiliar targets. In order to keep the avatar recognition at its highest, it is necessary to take the culture in the specific country into account. The reason is that different countries and cultures have different facial features, hairstyles or accessories needed to create a distinguishable avatar, that provide enough information for the user to recognize.

These results shows that there is a basis for using voices together with customizable avatar faces in a contact book. The studies showed that avatar recognition has a high recall accuracy as a result of customization and familiarity. Furthermore the voice snippets showed to have a higher recall accuracy over time than avatar recognition, making it useful for disambiguation. It was also found that the quality of the voice snippets could be reduced significantly and still provide an above chance recall.

This goes well with illiterates incapability to read because the need for recognizing names or numbers are removed in a contact book relying on audio and visual features.

Due to various differences in laws around the world, the voice recording feature will encounter problems due to privacy laws.

Further work would include a better graphical representation in the application and a test would be performed on illiterates to see if the tactics they apply will effect the design of the interface. It would be interesting to see if adding text to the icons throughout the prototype as a secondary feature, would help the illiterates in the process of learning how to read. The results could also be applied to other areas, such as interfaces used by children or interfaces by public services that targets illiterates.

REFERENCES

1. J. Chipchase, “Understanding non-literacy as a barrier to mobile phone communication,” *Blue Sky*, 2005. 1, 2

2. Z. Lalji and J. Good, “Designing new technologies for illiterate populations: A study in mobile phone interface

design," *Interacting with Computers*, vol. 20, pp. 574–586, 2008. 1, 2

3. E. Eme, A. Lacroix, and Y. Almecija, "Oral narrative skills in french adults who are functionally illiterate: Linguistic features and discourse organization," *Journal of Speech, Language, and Hearing Research*, vol. 53, pp. 1349–1371, October 2010. 1
4. A. Bhamidipaty and D. P., "Symab: Symbol-based address book for the semi-literate mobile user," *INTERACT*, pp. 389–392, 2007. 1, 2
5. E. Friscira, H. Knoche, and J. Huang, "Getting in touch with text: Designing a mobile phone application for illiterate users to harness sms," *DEV*, March 2012. 1, 2
6. A. Joshi, N. Welankar, N. BL, K. Kanitkar, and R. Sheikh, "Rangoli: A visual phonebook for low-literate users," *MobileHCI*, September 2008. 1, 2, 4
7. H. Knoche and J. Huang, "Text is not the enemy: How illiterates use their mobile phones," *CHI*, May 2012. 2
8. I. Medhi, S. Patnaik, E. Brunskill, S. N. Gautama, W. Thies, and K. Toyama, "Designing mobile interfaces for novice and low-literacy users," *Transactions on Computer-Human Interaction*, vol. 18, April 2012. Article 2. 2
9. S. I. Ahmed, M. Zaber, and S. Guha, "Usage of the memory of mobile phones by illiterate people," *DEV'13*, 2013. 2
10. B. M. Chaudry, K. A. Siek, and J. L. Welch, "Mobile interface design for low-literacy populations," *IHI*, January 2012. 2
11. C. C.-F. Or and H. R. Wilson, "Face recognition: Are viewpoint and identity processed after face detection?," *Elsevier Vision Research*, 2009. 2
12. Y. Sun, X. Gao, and S. Han, "Sex differences in face gender recognition: An event-related potential study," *Brain Research*, vol. 1327, pp. 69–76, 2010. 2
13. S. G. Young, K. Hugenberg, M. J. Bernstein, and D. F. Sacco, "Perception and motivation in face recognition: A critical review of theories of the cross-race effect," *Personality and Social Psychology Review*, vol. 16, no. 2, pp. 116–142, 2012. 2
14. M. A. Webster and D. I. A. MacLeod, "Visual adaptation and face perception," *Phil. Trans. R. Soc. B*, vol. 366, 2011. 2
15. P. Elftmann, "Secure alternatives to password-based authentication mechanisms," 2006. Diploma Thesis, Pages 23-26. 3
16. R. G. F. Jr. and R. B. A. Jr., "What makes a face memorable? the relationship between face memory and emotional state reasoning," *Personality and Individual Differences*, vol. 49, pp. 8–12, 2010. 3
17. O. C. and B. V., "Familiarisation with faces selectively enhances sensitivity to changes made to the eyes.," *Perception*, vol. 30, no. 6, pp. 755–764, 2001. 3, 8
18. G. V. Belle, P. D. Graef, K. Verfaillie, T. Busigny, and B. Rossion, "Whole not hole: Expert face recognition requires holistic perception," *Neuropsychologia*, vol. 48, pp. 2620–2629, 2010. 3, 7, 9
19. R. Wang, J. Li, H. Fang, M. Tian, and J. Liu, "Individual differences in holistic processing predict face recognition ability," *Psychological Science*, vol. 23, no. 169, 2012. 3
20. M. L. Gavrilova and R. V. Yampolskiy, "Applying biometric principles to avatar recognition," *International Conference on Cyberworlds*, 2010. 3
21. S. Berg, A. S. Taylor, and R. Harper, "Mobile phones for the next generation: Device designs for teenagers," *CHI*, vol. 5, no. 1, pp. 433–440, 2003. 3
22. S. Campanella and P. Belin, "Integrating face and voice in person perception," *Trends in cognitive science*, vol. 11, no. 12, 2007. 3
23. F. Joassin, M. Pesenti, P. Maurage, E. Verreckt, R. Bruyer, and S. Campanella, "Cross-modal interactions between human faces and voices involved in person recognition," *Elsevier Cortex*, vol. 47, pp. 367–376, 2011. 3, 4
24. K. Lander, H. Hill, M. Kamachi, and E. Vatikiotis-Bateson, "It's not what you say but the way you say it: Matching faces and voices," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 33, no. 4, pp. 905–914, 2007. 3
25. P. K. Kuhl, "Who's talking?," *Science AAAS*, vol. 333, no. 529, pp. 367–376, 2011. 3
26. B. R. Clifford, "Voice identification by human listeners: On earwitness reliability," *Law and Human Behavior*, vol. 4, no. 4, 1980. 3
27. F. McGehee, "The reliability of the identification of the human voice," *The Journal of General Psychology*, vol. 17, no. 2, pp. 249–271, 1937. 3
28. R. S. Newman and S. Evers, "The effect of talker familiarity on stream segregation," *Journal of Phonetics*, vol. 35, pp. 85–103, 2007. 3
29. A. D. Yarmey, A. L. Yarmey, M. J. Yarmey, and L. Parliament, "Commonsense beliefs and the identification of familiar voices," *Applied Cognitive Psychology*, vol. 15, pp. 283–299, 2001. John Wiley and Sons. 3
30. Z. Zvonar and R. Baines, "Integrated solutions for gsm terminals," *Journal of Wireless Information Networks*, vol. 3, no. 3, pp. 147–161, 1996. 3
31. A. M. Kondoz, *Digital Speech: Coding for Low Bit Rate Communication Systems*. John Wiley and Sons Ltd, second ed., 2004. Pages 1-3. 3, 4, 5
32. T. U. S.r.l., "Face your manga." `http://faceyourmanga.com`, 2008. [Online; Last accessed 16-May-2013]. 4