It’s Time to Personalise the Voice:

Designing for the Voice in Voice Technology

Dónal Kearns

A research paper submitted to the University of Dublin,
in partial fulfillment of the requirements for the degree of
Master of Science Interactive Digital Media

2020

Declaration

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at: http://www.tcd.ie/calendar

I have also completed the Online Tutorial on avoiding plagiarism 'Ready, Steady, Write', located at http://tcd-ie.libguides.com/plagiarism/ready-steady-write

I declare that the work described in this research Paper is, except where otherwise stated, entirely my own work and has not been submitted as an exercise for a degree at this or any other university.

Signed: _____

Dónal Kearns

8th May 2020

Permission to lend and/or copy

I agree that Trinity College Library may lend or copy this Research Paper upon request.

Signed: _____

Dónal Kearns

8th May 2020

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor Jack Cawley for his valuable input throughout this research. Not only was he supportive of my ideas but also gave me the confidence to complete this research and for this, I am extremely grateful.

Secondly, I would like to extend my deepest appreciation to my parents Paul and Kathryn for their unwavering support, motivation and guidance throughout my third level education. Without them, this research would not have been possible.

Lastly, I would like to acknowledge my girlfriend Emily for her continuous support and encouragement especially through times I found particularly difficult and to the amazing friends I have made at Trinity College Dublin for enriching my college experience.

Abstract

Speech is the primary way of communicating with a human. The past six decades have seen a growing attraction towards developing technology to talk to humans. Consumers have been interacting with speech technology since the 1990s, from automated phone calls to mobile apps run primarily through voice assistants. With the advancements in artificial intelligence and technology, Voice User Interfaces are becoming globally accessible. Many users have tried voice recognition on their mobile phone, after it fails however, they never try the system again. The key to user centred design is to understand why this is the case. The primary concern of this research however is to explore the voice in voice technology.

When a device talks, what should its voice sound like? Most voice assistants share the same characteristic; a one voice fits all approach. A polite and friendly female voice being the most common. This research will demonstrate the social consequences of designing a voice and how personality, gender and language can have a major effect on how users perceive the technology. Voice alone is bounteous in social information. Voice sounds can convey various signals that humans can pick up naturally to identify, such as personality, age and gender of a talking human. The same applies when humans talk to computers, as the “Media Equation” points out that humans treat computers as real people and can have real social relationships (Reeves and Nass 1996).

Voice Cloning has recently emerged in the past year and may be on its way to being implemented into consumer devices and homes. Voice cloning is a deep-learning algorithm which can record a human voice and synthesize it to match the original voice. LyreBird, Resemble AI and iSpeech Deepsync are some applications using the voice cloning algorithm. Some applications can now mimic a human’s voice within only a few minutes of recordings, which improves naturalness and similarity. As my research has pointed out, while voice cloning can create a more personalised and natural experience for the user, voice identity can be a very sensitive topic and can cause serious problems if not developed correctly. If the technology is not designed correctly; it can lead to vulnerable users who will not be confident in using the technology again. More user-centred research needs to be done to truly understand how users will react to this technology.

Table of Contents

List of Figures

List of Abbreviations

UI	User Interface
AI	Artificial Intelligence
VUI	Voice User Interface
IVR	Interactive Voice Response
HCI	Human Computer Interaction
SDS	Spoken Dialogue System
CUI	Conversational User Interface
GUI	Graphical User Interface
OVE	Orator Verbis Electric
LPC	Linear Prediction Coding
VRCP	Voice Recognition Call Processing
SALT	Speech Application Language Tags
VoiceXML	Voice Extensible Markup Language
ASR	Automatic Speech Recognition
DNN	Deep Neural Network
NLP	Natural Language Processing

1. Introduction

1.1 Context

There have been two major eras in Voice User Interfaces (VUIs). The first era consisted of interactive voice response (IVR) systems. These systems were capable of recognising human speech over the phone to carry out tasks. The emergence of applications such as Google Assistant, Siri and Cortana, which combine auditory and visual information, and voice-only devices, such as Google Home Assistant and the Amazon Echo, were the beginning of the second era (Pearl 2016).

The idea of having a conversation with a computer seemed futuristic not very long ago. However, in recent years, the technology has become widespread and inexpensive bringing voice technology into people's everyday life. However, this has led to privacy and security concerns for those using these devices in their homes. The acceptance of voice assistants depends on a number of factors. One of the factors is the voice, which this paper will address. Today, voice assistants have a one voice fits all approach, encouraging gender binary, social stigma, and pollyanna. The problem is that these devices were not designed for the user.

As technology improves, computers will soon be able to have a conversation with humans. In academia, the goal is to produce a Spoken Dialogue System. This is a term used in Human Computer Interaction (HCI) to describe voice technology that perfectly satisfies the user's experience. The requirements for Spoken Dialogue Systems are: user satisfaction, quality of service and customisation. The primary focus of this research is to look further into voice customisation.

1.2 Motivation

There are many advantages to using voice technology. It enables users to ask simple commands from dialling a phone number, playing music to asking for simple information. It has become a very popular user interaction (UI) channel, especially in smart watches, home appliances and motor vehicles (Feng 2017). Google's Assistant, Microsoft's Cortana, Amazon's Alexa and Apple's Siri are the most popular assistants today, which are embedded into the smart phone (Hoy 2018). These assistants will continuously improve in response time, knowledge, aspects

of human intelligence and less errors. The relationship between the users and the technology will change as it will understand the user in a more complex and deeper way. Virtual assistants could become users councillors and even friends. We are a long way from seeing this happening. Although Spoken Dialogue Systems are currently the closest available to matching the ultimate goal of voice technology, they are a long way from reaching its full potential. When designing for Spoken Dialogue Systems, the user must come first if we want to achieve this.

1.3 Research Question/Objective

Recently there has been a lot of research on the natural conversation experience. However, there are still a lot of unanswered questions and not a lack of testing done in this field. For example, future voice user interfaces may be able to clone the user's voice. This realistic voice and possible lack of human characteristics, such as breathing or sighing, may cause people's perception of the technology capabilities to be unsettling (Murad et al. 2019). Another aspect for a potential feature could be instant voice responses. For example, virtual assistants may eventually have the potential to respond faster with greater accuracy than humans. However, this could lead to a mistrust in the technology and a fear of it listening to their conversations.

As spoken dialogue systems are currently being developed, it is difficult to test on users without a fully developed prototype. A method which can help developers and engineers determine what the users want in development mode, is a Wizard-of-Oz experiment. This method allows the users to interact with an application which they perceive as autonomous, but in reality, it is a human operating the application in disguise (Fraser and Gilbert 1991). The technique helps determine the understanding of how a user would react to a fully working product. The goal of this research is to ask; "Firstly, does the use of some personalised human characteristics improve the user's experience, and secondly, does complete voice cloning further improve this experience?"

2. State of the Art

2.1 History of Mimicking Speech

There have been attempts to develop technology that mimic human's speech communication since the latter half of the 18th century. The early interest was on creating speaking machines, rather than recognising and understanding the human's speech (Juang and Rabiner 2005). Christian Kratzenstein produced vowel sounds using resonance tubes and organ pipes in 1779. (whetty 1999). Later in 1971, Wolfgang von Kempelen invented the Acoustic-Mechanical Speech Machine. This was able to produce single sounds and sound combinations (Dudley and Tarnoczy 1950). In this mid-1800s, Charles Wheatstone further implemented von Kempelen's speaking machine. Most of the consonant words and some sound combinations and even full words were now then possible to produce (Wheatstone 1879). The desire for automation of simple tasks has been around since 1881, when Alexander Bell, Chichester Bell and Charles Tainter invented a recording device to respond to incoming sound pressure. Grooves were cut by a stylus onto a rotating cylinder coated in wax (similar to a vinyl record). This led to the formation of Volta Graphophone Co. in 1888 which led to the manufacturing of recording machines which reproduced sound in offices (Juang and Rabiner 2005). It was later trademarked as "Dictaphone" in 1907 by Columbia Graphophone Co. The Phonograph was invented at a similar time by Thomas Edison. These products were invented to record letters and notes for secretaries, allowing them to type out the recordings on a typewriter later (Juang and Rabiner 2005).

2.2 History of Speech Synthesis

In 1922 the first electrical synthesis device was developed by Stewart (Klatt 1987). Consisting of a buzzer and two resonant circuits, the device managed to generate single static vowels. However, it could not generate consonants or connected utterances. Obata and Teshima in 1932 discovered the third format in vowels in Japan (Schroeder 1993). The VODER (Voice Operating Demonstrator) is considered as the first speech synthesizer device (Klatt 1987). The VODER was an electrical equivalent of the Wheatstone's mechanical speaking machine. It consisted of a wrist bar for selecting voice or noise sources and a foot pedal to control the frequency. The user's fingers controlled the output levels. Ten bandpass filters altered the

frequency range. The VODER required a lot of skill from the user to be able to play but in terms of the evolution of speaking machines, it was considered an important milestone (Dudley 1940). This was the first time that speech could be produced artificially.

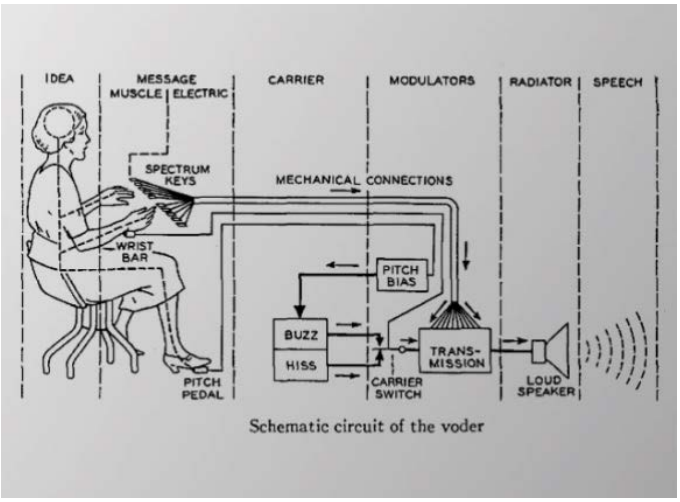


Figure 1. The VODER (Dudley 1940)

PAT (Parametric Artificial Talker), was the first formant synthesizer (Klatt 1987). Developed by Walter Lawrence, it consisted of three parallel electronic formant resonators. A buzzing sound was the input signal. The fundamental frequency, voice amplitude, and three formant frequencies were controlled by a moving glass slide. This converted paint patterns into functions. The Orator Verbis Electric (OVE I) was introduced at around the same time as the PAT. This however had formant resonators connected in series. It could only produce vowel-like sounds. Over the next ten years from 1952, there was big improvements in control strategies and synthesizers, reaching closer to human sentences. In 1962, the OVE II synthesizer was developed, also by Gunnar Fant, which improved by having a separate static branch to model the transfer function of the nasals, vocal tract for vowels, and obstruent consonants. (Klatt 1987).

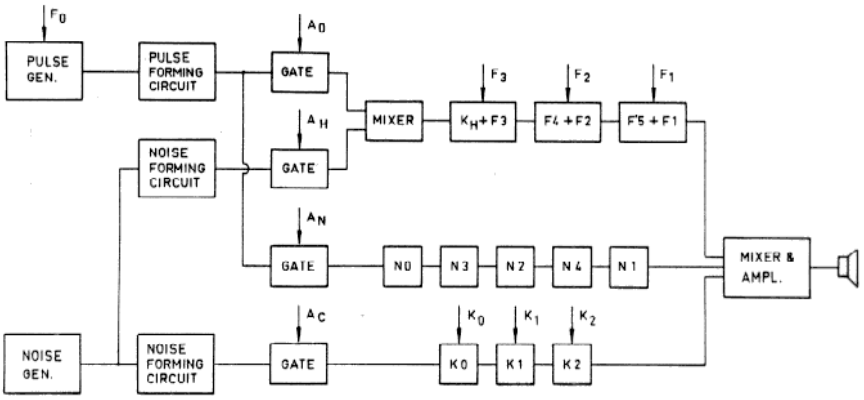


Figure 2. The OVE II Speech Synthesizer (Klatt 1987).

In 1972 John Holmes invented his parallel format synthesizer. He hand-tuned a sentence so well that some listeners could not tell the difference between the natural and synthesized voice (Klatt 1987). John introduced the parallel formant synthesizer with the Joint Speech Research Unit. It could generate “speech signals up to a maximum frequency of 4kHz”. (Holmes et al. 1990). Electrotechnical Laboratory in Japan developed the first full text-to-speech system in English (Klatt 1987). Developed by Noriko Umeda, it had a syntactic analysis module with sophisticated heuristics. However, the quality is nowhere near to present systems as the speech was monotonous. The MITalk text-to-speech system was later developed in 1979 in M.I.T labs (Allen et al. 1987). The system was used in Telesensory Systems Inc. (TSI). Dennis Klatt introduced his Klattalk system in 1981 (Klatt 1987). Both the Klattalk and MITalk systems inspired many synthesis systems designed today.

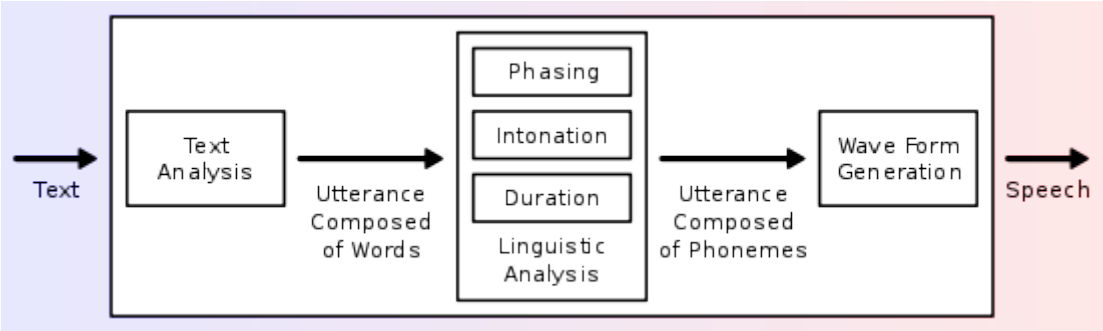


Figure 3. Text to Speech (TTS) Overview (Allen et al. 1987)

Kurzweil introduced the first reading aid in 1976. With an optical scanner, the reading machines were able to read multiple font written text for the blind. The system was mainly used in service centres and libraries as it was costly for the public (Klatt 1987).

The Votrax chip was the first integrated circuit for speech synthesis. The chip included a simple low pass smoothing circuit and a cascade formant synthesizer. A Votrax-based Type-n-Talk system was introduced in 1978 by Richard Gagnon (Klatt 1987). Texas Instruments two years later introduced a Speak-n-Spell synthesizer based on linear prediction coding (LPC). The system was a success as it was used as an electronic reading aid for children. The Prose-2000 was introduced by Speech Plus Inc. This was developed in 1982, around the same time as the Echo low-cost diphone synthesizer, which was developed by Street Electronics. In 1983, DECtalk and Infovox SA-101 were introduced as the first commercial versions. Sufficient power and flexibility to plug in improved versions was included in the DECtalk hardware (Klatt 1987).

2.3 Text-to-speech

Speech-to-text process was regarded as the first step in enabling machines to understand and correctly respond to human speech. In the 1990's, call centres started to emerge which handled telephone calls from customers. To reduce the cost, automatic speech recognition technology was used to automate the calls. In 1992, AT&T was one of the companies that introduced Voice Recognition Call Processing (VRCP) to route and handle calls. (Juang and Rabiner 2005). Jupiter and Pegasus were noteworthy systems that used VRCP, which were developed by Victor Zue in Massachusetts Institute of Technology. Pegasus was a spoken language interface for air travel planning. It enabled users to book flights with American Airlines (Zue et al. 1994). In 1997, Jupiter was developed, which was a conversational interface that allowed users to receive weather forecast information using spoken dialogue with their telephone (Zue et al. 2000). The goal of these machines was communication, rather than recognising the user's words (Juang and Rabiner 2005).

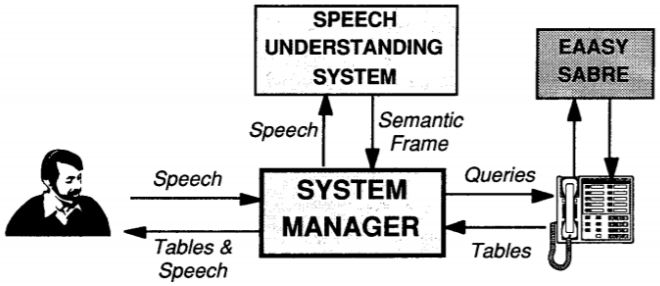


Figure 4. Pegasus travel planning system (Zue et al. 1994).

These systems were the beginning of the first era of VUIs. They were known as Interactive Voice Response systems (IVR), as they could understand human speech over the telephone (Pearl 2016). They became popular in the 2000s as they allowed anyone with a telephone to book flights, transfer money or hear traffic information. Many of these IVR systems were more conversational than current VUIs. This was because IVRs kept track of what information the user said until the call had ended. Some IVR systems, like the Charles Schwab's trading service, surprisingly received a lot of traffic. The companies realised that some customers were much happier to call into an automated system rather than a real person, as they could use the system an unlimited amount of times without bothering a member of staff. It also allowed customers to do tasks at any time, due to the 24/7 nature of IVR systems (Pearl 2016).

2.4 Voice on the web

Until the early 2000s, interactive speech applications required APIs (Application Programming Interfaces). Speech Application Language Tags (SALT) and Voice Extensible Markup Language (VoiceXML) emerged, which allowed developers to use existing web infrastructure and build on it, using the standard internet protocols (McTear 2004). VoiceXML, published in 2000, is an open standard markup language for voice applications on the web. It was developed for HTML to make it easy for web developers to create voice applications for the web (VoiceXML 2000). SALT is used in HTML pages to add voice recognition to web applications. It enabled multimodal access which allowed users to interact with web applications in several ways. One of SALT's purposes was to encourage designers to focus on core user interface design issues, rather than software engineering and computer details.

2.5 Voice Assistants

Voice user interfaces experienced little changes in the coming years, and it was not until the 2010s when IVR systems started being replaced and upgraded. Voice-only devices, for example, smart speakers, and voice integrated mobile apps became mainstream and replaced IVR systems (Fernandes 2018). This was known as the "second era of VUIs" (Pearl 2016). Google Now, Siri and Cortana started to emerge on mobile phones and computers which combined visual and auditory information. Later came standalone devices, like the Google Home Assistant and Amazon Alexa, which interacted with voice only. These devices are all under the umbrella term of Voice Assistant as they are software agents that run on smartphones or purpose-built speakers (Hoy 2018). Siri, Apple's voice assistant was the first to emerge on the market in 2010, which was integrated into the iPhone 4S in 2011. Cortana followed shortly in 2013 on Microsoft Windows devices. Amazon released the first standalone voice assistant in 2014, called Alexa. Google followed in 2016 with its home speaker and mobile app for android phones. These voice assistants differ from earlier voice technology as they can respond to a much larger amount of commands, due to their constant connection to the internet. The user's voice command is sent to a central computing system to analyse it. The central computing system then provides the appropriate response (Hoy 2018). The development of machine learning methods, an increase in computing power, the availability of larger amounts of linguistic data and a better understanding of the structure of the human language in social

contexts have all been credit with this recent improvement in natural language processing (Hirschberg and Manning 2015).

Figure 5. (a) Apple HomePod (b) Amazon Echo (c) Microsoft Harman (d) Google Home (Kepuska and Bohouta 2018)

As people created more and more text online to be analysed, scientists have taken advantage of the information to train voice assistants to naturally listen and respond to requests in more meaningful ways. Unlike older voice technology, voice assistants can “parse requests”, phrased in different ways (Hoy 2018). This is called natural language processing.

Companies with voice assistants use different techniques to improve their technology. Amazon has advanced deep learning functionalities of Automatic Speech Recognition (ASR) that converts speech to text. Amazon provides Natural Language Understanding (NLU) to understand the text which enables its developers to design engaging user experience applications and realistic conversational interactions (Amazon 2017). The Google Assistant is improved by the Deep Neural Network method (DNN). Microsoft uses the Microsoft Azure Machine Learning Studio (Microsoft 2010). Facebook has recently launched Messenger M, which combines contextual memory with machine-learning algorithms. Messenger M is being trained with supervised learning, which is when the computer learns from what human trainers teach it (Kepuska and Bohouta 2018).

2.6 Spoken Dialogue Systems

As artificial intelligence is constantly improving, it is beginning to emerge in everyday applications such as healthcare, gaming and media. Spoken Dialogue Systems are beginning to emerge, which allows the user to partake with the computer using natural spoken language. This enables users to interact human-to-human conversation with software. Rather than the user giving voice commands to a computer, spoken dialogue systems allow the user to communicate in their natural language. Voice assistants such as Amazon Alexa and Google Assistant focus on short interactions, like answering simple questions or playing a requested song, rather than longer free-flowing conversations (Khatri et al. 2018). Longer and free-form voice conversations are often open domain. Topics change in natural conversations over time. There can be an unlimited amount of responses, even if the two interactors share similar interests or have similar backgrounds.

Speech was not used for input until the 1980s. (McTear 2004). ELIZA was the first system to simulate conversation in 1966 (Weizenbaum 1966). The conversations with ELIZA was impressive for its time, however the conversations were limited, and the system was only text-input. This system was called a simulated conversation as it simulated conversational interaction. Rather than using theories and models from natural language processing and AI, this approach used pattern matching (McTear 2004)

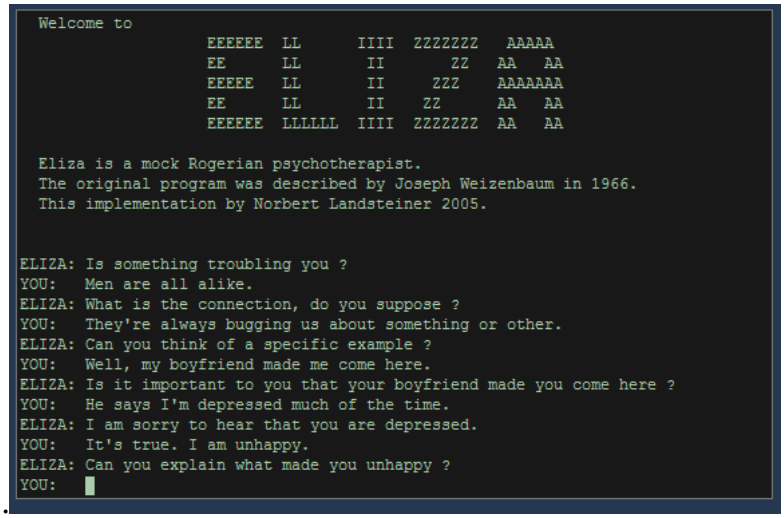


Figure 6. Eliza Simulated Conversational System (Weizenbaum 1966).

Conversational systems can be traced back to the 1950s (McTear 2004), however, with major advances in speech technology in the last decade, working systems have been developed. The

voice-command in voice assistants have limited capabilities in human conversation, however, they have demonstrated how the use of voice technologies can enhance the user experience (Lison and Meena 2014).

Some companies have started to attempt to implement spoken dialogue systems in their technologies. These systems have been seen in personal assistants, smart-home environments and tutoring systems (Lison and Meena 2014). For example, The Ford Model U Concept vehicle has a speech recogniser and touch screen. This allows the driver to navigate, request the weather forecast and request a phone call, all through voice recognition. It does this however using a natural language spoken dialog interface, rather than a command and control interface, found in current vehicles. This means that the system is conversational, reducing the memory burden for the end user (Pieraccini et al. 2003). Semio, developed at the University of Southern California, is a platform which allows users to talk to robots through body language and natural communication. Developers were able to create gesture/speech-based applications. Non-expert users were then able to access these robot applications through natural communication (Mead 2017). Nao is another interactive robot that is able to extract and speak Wikipedia content using multimodal interactions. WikiTalk is implemented inside the robot, which is a Spoken Dialogue System. This provides Nao with an unlimited range of topics for discussion. Nao also has face tracking, hand gestures, nodding features and allows tactile interruptions (Csapo et al. 2013).

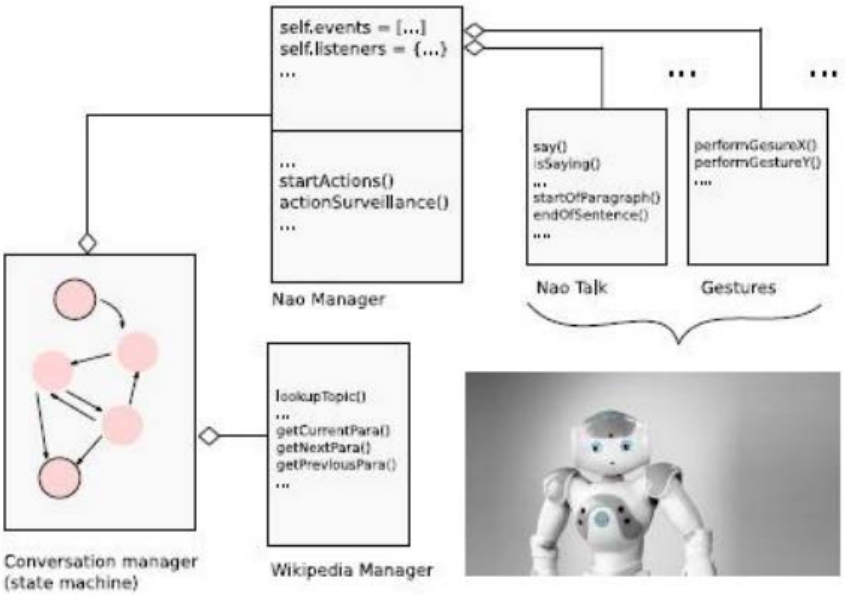


Figure 7. Overall view of the system architecture (Csapo et al. 2013).

From this research, it is obvious that we are not close to developing perfect Spoken Dialogue Systems. There are individual modules which process all of the information which are then passed further onto other modules in the system (Lison and Meena 2014). The closest system to a Spoken Dialogue System is the Furhat Social Companion Talking Head. An attention controller is installed to handle the dialogue content without being distracted by the devices capturing the input. It uses a projection system to render facial expressions, with a motor that can move the head (Al Moubayed et al. 2012). OpenDial is an open source toolkit for developers to build and evaluate spoken dialogue systems. The toolkit has already been developed in robots and car assistants. Developers can provide an XML-encoded domain to apply it to a dialogue application (Lison and Kennington 2015). There are many more smart devices, social robots and virtual agents that are being developed to be conversational. However, these devices do not go further beyond the command and control devices (McTear et al. 2016).

Developing Spoken Dialogue Systems is very difficult. Many errors are likely to emerge in the processing pipeline due to speech recognition errors which is caused by ambiguous words and accents. The cognitive operations must be completed in a suitable time, as dialogue is a real time process for humans. There are many components in a Spoken Dialogue System, including speech recognition, speech synthesis and language understanding. The image below is the current design for a Spoken Dialogue System. The current system is flawed because each module processes the information individually. This is typical in pipeline architectures; therefore, they are unable to utilize information from other modules to help with speech recognition (Lison and Meena 2014).

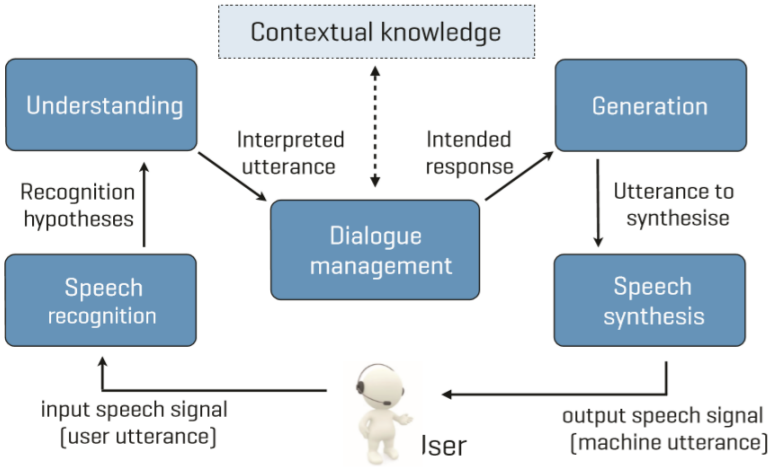


Figure 8. A spoken dialogue pipeline model architecture (Lison and Meena 2014).

2.7 Current problems with Voice Assistants

In recent years, voice technology has become much more conversational. CUIs have become the norm for everyday users. Therefore, the user's expectations of voice assistants have increased greatly. Users expect the technology to be flawless and to be able to match human capabilities in speech. Unfortunately, the technology is currently not sophisticated enough, increasing the gap between user experience and expectation of current conversational user interfaces (Luger and Sellen 2016). Despite the huge tech giants eager to invest in conversational user interfaces, there is still little knowledge of how the agents are used in everyday settings, in the field of HCI. Currently, we fail to understand "which factors influence acceptance and success in such scenarios" (Kopp et al. 2005).

Numerous user studies have revealed that users have several concerns about trusting these agents. Cowan et al. (2017) conducted a qualitative analysis on users who have used voice assistants but have chosen to not use them regularly. They found that social embarrassment and cultural norms are one of the reasons users may not trust the devices. Clark et al. (2019) have found in their research that some users needed a long-term bond and common ground before they can trust the technology. Nevertheless, it has been said that there has been little research in how to define trust, how to measure it, and how we can make the interfaces more trustworthy (Edwards and Sanoubari 2019). There can be a range of issues with fundamental differences. A user may not trust a voice assistant due to it not being able to complete a task or a user may think the voice assistant is maliciously trying to steal their data (Edwards and Sanoubari 2019). A common trust issue for users using CUIs is a lack of as to what data is being collected and how it is being treated. This notion has been explored in other mediums (Kiousis 2001) and may help in researching the trust issue in CUIs. This lack of trust resembles that of websites and online news in the early days of the world wide web (Flanagin and Metzger 2007).

To conclude, understanding user's lack of trust for voice technology can be very challenging and confusing, as defining trust can be heterogeneous, while the concept of trust can have an assorted amount of meaning for users (Edwards and Sanoubari 2019). To move forward in gaining user's trust, researchers and developers must look to previous examples in similar mediums, and most importantly continue to design for the user first.

2.8 The future of the conversational interface

Despite the growing problems with trust issues in voice technology, developers and engineers are continuing to make the technology more conversational and human, thanks to the increased development in artificial intelligence. Voice assistants will eventually be able to fully understand the users, but, do users want that? (Burbach et al. 2019). Despite what the users want, there are many reasons for the production of conversational agents. Some examples are: advances in artificial intelligence, improved processing computer power, increased connectivity, advances in natural language processing and major interests from the biggest technology companies (McTear 2016). As conversational artificial intelligence continues to increase in popularity, how "human-like" will conversational interface become? The answer to that question is very difficult to answer because the relationship with users and conversational user interfaces in the future will change. This will result in "new user behaviours as well as new social norms and user expectations" (Brandtzaeg and Følstad 2018). In 2016, IKEA's chatbot, Anna, retired due to unforeseen circumstances. The reason why IKEA shut Anna down was unclear, however customer feedback has given clear reasons why. It struggled with balancing robot and human aspects, often confusing the user and encouraging them to ask unrelatable questions. The assistant was regarded as "too human", and as Magnus Jean pointed out, if the assistant tries too hard to be natural, it can divert the user from the real purpose, which most of the time is giving the right answer as quickly as possible (Wakefield 2016). This example proves that it may not be the best solution to create voice assistants as realistic as humans. The only solution to finding out what users want is to continue testing on users to know more about how people experience interactive conversational user interfaces and to understand the user needs that motivate the future use of voice technology (Brandtzaeg and Følstad 2018).

Figure 9. Anna, IKEA's virtual assistant (Wakefield 2016)

2.9 Voice Cloning Text-To-Speech Synthesis

Tacatron, developed by Google, is a system that uses text-to-speech synthesis that generates human-like speech from a user's text input. The neural networks are trained by speech examples and text transcripts only. The code is open source, allowing anyone to train the model completely to create a unique voice. Tacatron 2 is the latest version. The voice samples sound very realistic according to user studies, however, the system has difficulties pronouncing complex words and fails to generate audio in real time (Wang et al. 2017). Adobe VoCo is a text-based insertion and audio replacement tool, currently unreleased to the public. Developed by Zeyu Jin, the system can synthesize new words or sentences that can blend seamlessly into an existing narration. The output is often almost identical from the original narration and is indistinguishable to humans (Jin et al. 2017). It can mimic a human's voice with a 20-minute voice recording.

Lyrebird AI is a voice synthesis system that can create a digital voice of a human from a one-minute voice recording. Lyrebird allows developers without any knowledge of VoiceXML (used for developing voice response applications) to use the application, thanks to its simple graphical user interface (GUI), which reduces a significant amount of time (Lyrebird 2017). These examples prove that the technology is not far away from computers matching human voices. Voice synthesis systems will be so fast at mimicking human voices, voice assistants will be able personalise its voice to match the owner's voice. Current voice synthesis systems are so advanced that researchers are already trying to measure, using tools, on how to detect a synthesised voice from a human voice (AlBadawy et al. 2019).

3. Methodology

The purpose of this research is to look at how the quality of the voice used in voice technology can be improved to better the users' experience. The previous chapter has reviewed the history of voice technology and how it has evolved to its present format. However, aforementioned, making the voice of the user interface as realistic as possible may not be the best option. Rather than researching on how to improve the current voice technologies this paper will instead focus on preventing future voice technologies from causing problems that may reduce the user's experience even more. Marketers always want the greatest and latest features in interfaces while designers want to make sure that the latest features will improve the user's experience for the user. Designers would rather develop a balance between usability and technology (Nass and Brave 2005). This is the current problem facing voice assistants today.

3.1 Wizard-of-Oz Method

Burbach et al. (2019) conducted a choice-based analysis to find out the acceptance relevant factors of voice assistants. The study found that users rated natural language more than rating it negatively. However, the scientists asked the question directly, rather than testing with an actual natural voice. It is recommended therefore to test this using the Wizard-of-Oz experiment (Burbach et al. 2019). A Wizard-of-Oz experiment is when users interact with an application which they believe is autonomous, but it is a human (wizard) operating the application in disguise (Fraser and Gilbert 1991). The technique helps determine the understanding of how a user would react to a fully working system. The study collects proactive and situated data before the system is fully developed. It is a great method for researchers as it is easy and inexpensive to set up, compared with fully building a completed system. The data revealed can discover phenomena of interest, linguistic and behavioural phenomena and can create a solid foundation for the next prototype (Oviatt and Adams 2000).

A Wizard-of-Oz method can be a good way to test users on whether they will prefer having a voice cloning tool in a voice assistant. The idea would be to implement one of the current voice cloning software into a physical voice assistant. A hidden wire will extend from the voice assistant to a laptop, connecting it to an external speaker. The Wizard will then input text via the laptop which will play out of the external speaker inside the voice assistant. In order to

clone the participants voice, they will need their voice recorded for at least a minute for the voice cloning software to gather the necessary information. The user will ask the assistant questions, thinking it is a working prototype, while the Wizard inputs the answers on the keyboard. This is one approach to finding out user's reactions accurately. However, this may lead to potential errors, as the response from the voice assistants may seem unnatural and slow to respond. While some listeners may prefer a slow response to digest the information, other listeners may not.

3.2 Is realism and accuracy always better?

As discussed in the previous chapter, research has shown that making technology more realistic is not always the best option. Morishima Mori (1970) coined the term 'The Uncanny Valley' to describe when robots appear humanlike, they become more appealing to the user, however only up to a certain point. Once the robot reaches that threshold, the user starts to exhibit negative feelings towards it. Positive feelings start to emerge again when the user can distinguish that the interface is not an actual person.

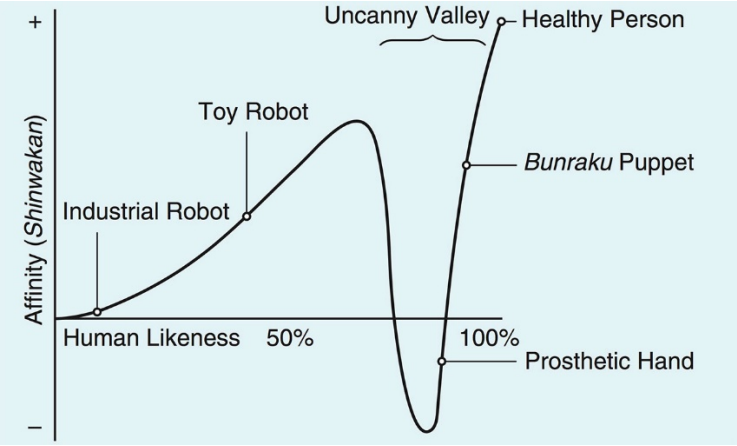


Figure 10. The Uncanny Valley (Mori 1970).

As Nass and Brave (2005) pointed out, more humanlike and realistic conversational interfaces are not always better than a less realistic interface. An important point to note is that humanness maximised in one dimension, but not along other dimensions, is a big problem (Stern et al. 2006). Another point of note is that synthetic voices that use the word "I" receive more of a negative response than when a human voice says it (Nass and Brave 2005).

Another design principle that Nass and Brave (2005) noted is that accuracy is not always the best option. According to research, positive comments lead users to a more enjoyable experience, for example, “You have made a great choice”. The effects of flattery from robots are the same as the effects from humans (Nass and Fogg 1997). Developers sometimes do not always realise this, as computers are built for quality and precision while society is built on “honesty is not always the policy”. According to Nass and Brave (2005), accuracy can burden users.

Some human behaviour is preferred and sometimes expected in voice technology. For example, Bickmore and Cassell (2005) developed an agent which generated small talk at the start of a conversation with a human, typical in human-human conversations. Users were tested interacting with an agent (voice technology) that spoke task-oriented interactions and an agent that generated additional social dialogue at the beginning. It turned out that the latter choice was the preferred option as it was more human-human, which increased trust. Interestingly, the results showed that the latter also required less cognitive load for the user, compared to the task-oriented agent (Bickmore and Cassell 2005). ‘The Uncanny Valley’ proves that having a realistic voice may improve the user’s experience, as the user will know he/she is talking to a computer from other dimensions. The realistic voice also may seem to appear less accurate than a synthesised voice.

3.3 Should the Interface say “I”?

There has been plenty of research into the use of “I” in robotic speech, including a telephone-based action experiment (Huang et al. 2001). Sixty-four native English-speaking college students participated in the experiment. The group was randomly split into two, one half used a system with a synthetic voice while the other half used a system with a recorded voice. Half of each group were presented with a system that used “I” in the sentences while the other half did not. Speech Interface research have previously argued for using human-like-sounding speech as much as possible. However, present research has noted that it is more complicated than predicted. Rather than users preferring machine-like or human-like interfaces, Huang et al. (2001) points out that consistency is more important between content and voice. Consistency can have a big impact on the quality of the interface and the user’s behaviour. Therefore, human-like voices should be paired with human-like scripts and vice versa. This is because

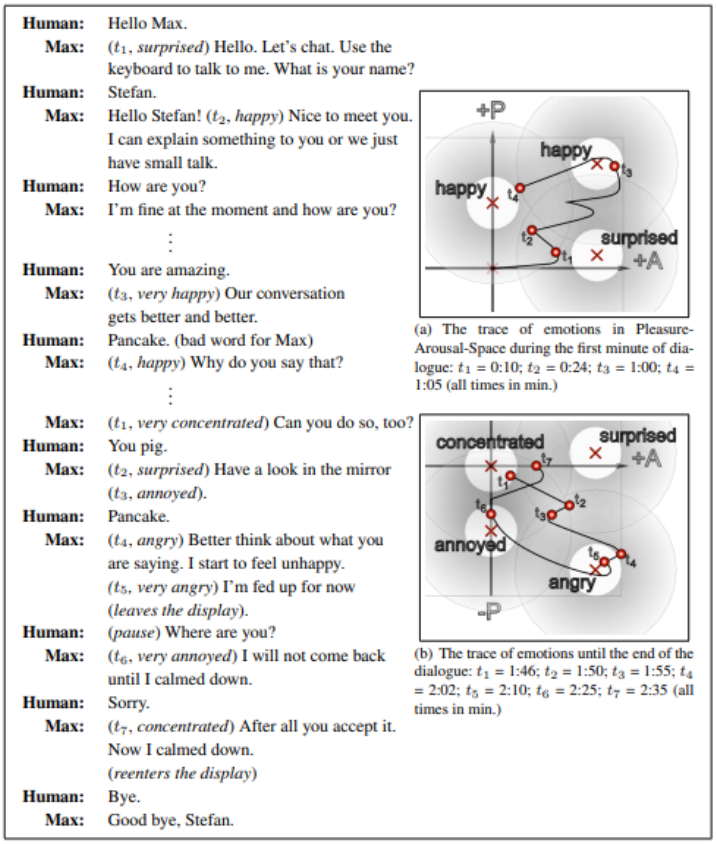
research has shown that humans do not like ambiguous categories. The results showed that users perceived a passive voice and synthesised speech less human than first-person and recorded speech.

A very important design point Huang et al. (2001) mentioned was that humanness or socialness in not always the preference for users. When designing a voice interface, the consistency effect, as well as the strong aversion towards human computers should be accounted for, rather than assuming that the more human-like the interface is, the more users will like it. When will the quality of speech be good enough to be perceived as a human? If the speech can mimic clarity, emotion and prosody of the human speech while at the same time retaining some indication that it is still machine-generated, the consistency effect and 'The Uncanny Valley' may disappear (Huang et al. 2001).

3.4 Effects of humour, fear and emotion in Voice Technology

Theorist in Human Computer Interaction and software engineers in the past have disparaged the idea of using humour (Nass and Brave 2005). Non offensive humour has been proved to be beneficial, to facilitate work, bond employees together in the workforce, improve socialisation and boost morale (Clouse and Spurgeon 1995). In voice interface design, humour has been underused, mostly due to the fact that humour can have major downsides if not used properly. Female voices with a sense of humour, in particular, can be perceived as sarcastic and aggressive (Nass and Brave 2005). Sexual, racial or ethnic humour should be avoided as it can be very offensive to certain users. Nass and Brave mention that one type of humour that has had a consistent positive effect on users is innocent humour. This is humour that is soothing and light. While a voice interface is ideal for implementing humour into interactions, designers must remember the principle of consistency. Jokes are not translated the same way prompts are translated and a certain type of humour is associated with certain personalities. Designers must be conscious when humour can be used and cannot be used. For example, when using banking or financial applications, humour may suggest an overly relaxed approach, while using humour when users are buying books, toys or music is reasonable as this is natural in real circumstances (Nass and Brave 2005).

Kostov and Fuduka (2000) predict there will be a surge in robotics with emotion. Therefore, personalisation in future voice user interfaces is a must for the complete user satisfaction. Results show that "voice emotion sensitive agents are feasible" (Kostov and Fukuda 2000). For users to feel that voice user interfaces are realistic, emotion is crucial. If designers want users to take a synthesized voice seriously, adding an emotion aspect is one of the first steps in achieving that. Even a limited and basic language processing computer program like ELIZA, will communicate more efficiently and effectively with users if the programme can express and perceive emotions (Picard 2000). The main reason designers should integrate emotion into conversational interfaces is because emotion increases the believability of the interface. Emotion also allows the interface to generate and respond in an appropriate manner, where deliberate and instant responses are inappropriate (Becker et al. 2007). According to Nass and Brave (2005), emotion decreases risk taking, regardless of how the user is feeling themselves. The voice assistant will therefore be judged more positively when a recommendation is given in an emotional voice (Nass and Brave 2005). Research has shown that integrating humour has had a positive effect on users. For example, Becker et al. (2007) proved that having a constant and positive voice has led to a more pleasurable experience (See figure 11).



(a) The trace of emotions in Pleasure-Arousal-Space during the first minute of dialogue: t_1 = 0:10; t_2 = 0:24; t_3 = 1:00; t_4 = 1:05 (all times in min.)

(b) The trace of emotions until the end of the dialogue: t_1 = 1:46; t_2 = 1:50; t_3 = 1:55; t_4 = 2:02; t_5 = 2:10; t_6 = 2:25; t_7 = 2:35 (all times in min.)

Figure 11. In this scenario, dominance is always constant and positive. The images show Max's emotions in the Pleasure-Arousal-plane (Becker et al. 2007)

On the other hand, should negative emotions be integrated into Conversational User Interfaces (CUI)? If so, should all or some negative motions be integrated? This is an idea which needs further researching and testing. An argument against this is that CUIs are built to be of help and why should they be allowed to be in a bad temper or have a mental breakdown? Humans however would find it irritating if the interface were unable to express "such an emotional state properly" (Becker et al. 2007). It would also be inappropriate if the full spectrum of emotions were limited. Therefore, negative emotion should be integrated, for believability, if it does not interfere with the performance or obstruct the accuracy of the technology. Emotion is crucial for human communication. It is the most powerful type of state to predict how a user will behave. So powerful, a big part of the brain used in emotion determines whether an image is a human or not (Nass and Brave 2005). Emotion is experienced in all daily activities, from sending a text message to driving to the shop. Emotion plays a critical role in every activity which involves a goal.

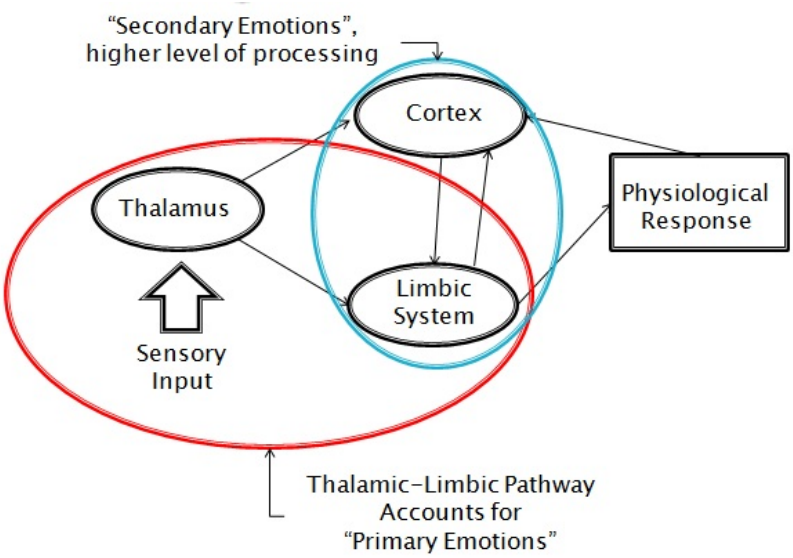


Figure 12. Neurological structure of emotion (Nass and Brave 2005).

The figure above represents three key areas of the brain – the cortex, the limbic system and the thalamus. The thalamus sends information to the cortex, which is used for higher-processing attention, and to the limbic system, which evaluates the information and the user's goals. If the information and the user's goals match the limbic system messages are sent to different parts of the body e.g. the heart is pumped faster when excited. The thalamus is also called the seat of emotion (Nass and Brave 2005). The primary emotions (circled in red above) are the body's first response and are easy to identify. They include sadness, happiness, anger and fear. Voice

user interfaces with eerily and unexpected sounds can activate primary emotions, and not always the good emotions. Secondary emotions are emotional reactions users have to other emotions. They include pride, frustration, shame or anxiety. These emotions are much more complex than primary emotions. Secondary emotions play most of the role in the design of VUIs (Nass and Brave 2005). Hence, these emotions need to be considered substantially more than primary emotions when designing. The interface should be able to express the full range of affective states, including aspects for communication, like attitude interpersonal stance and mood (Schröder et al. 2010). When the user reflects or expresses affective states, the technology should do the same. Furthermore, the voice assistant should be able to match the subtle changes in voice quality, as well as hesitations and intonation. Strictly speaking, it should be possible for the user to change the assistant's voice and personalise it to their needs. This will allow the user to pick and choose voice features that they like and discard the emotions they don't want in a voice assistant (Schröder et al. 2010). Some researchers believe most emotions are innate, while others believe most emotions are socially constructed (Nass and Brave 2005). People believe the latter is true as the limbic system can operate in an on or off manner. From this piece of evidence, emotions vary across different cultures considerably, with consistency between the cultures coming from social structure and not biology (Nass and Brave 2005). Therefore, it is vital to allow the user to personalise the assistant to suit their personal preference with emotion and feelings. Theorists have noted that there are six basic emotions, happiness, fear, sadness, joy, disgust and anger, common to all humanity, no matter what background or culture they come from (Ekman 1992). Therefore, it may be practical for voice assistants to have a default setting to include the basic emotions. If a user does not like the default setting, they should be able to personalise the assistant to suit themselves.

When adding emotion, it is important to get the pitch of the voice right. Emotion can be interpreted through amplitude, pitch range and rhythm (Ball and Breese 2000). A voice with fear, joy or anger will be faster and more accurate with a higher frequency, while a sad voice will be lower in pitch and slower. Users are very good at indicating emotion in computer generated voices. Users are around 60% accurate with a human voice and 50% with generated voices from computers (Nass and Brave 2005). However, the emotion in the voice needs to be engineered correctly to prevent the user from being confused.

3.5 Personality

Personality in voice assistants can be very powerful. It has the ability to influence a user's judgement of the assistant. It can also predict if they will trust and like it, as well as buy the product. The pitch, pitch range, speed and volume are very important on how the user will behave. These are the four aspects of voices that indicate personality (Nass and Brave 2005). The appropriate personality to give a voice assistant is simple; it must have every type and it should match that of the user speaking to it. This is called similarity attraction (Nass and Lee 2001). People who interact with others with similar personality have a more positive experience than with someone with a very different personality. Humans think that similar personalities in other humans are friendly, trustworthy and intelligent. Opposites attract because of personality similarity too. A person can like another person with an opposite personality even more if their personality changes to match (Moon and Nass 1996).

Human personality matches personality in synthetic voices. Research has shown that users have no problem when indicating if a voice is introverted or extroverted in a synthesised voice. The same research also showed that introverts preferred an introverted voice while extroverts preferred an extrovert voice (Nass and Lee 2001). This means that users can predict a voice assistant's personality, recorded or synthesised, and will have a preference that matches their personality. Nass and Lee (2001) proved that a voice with similar personality encouraged users to buy a product in an experiment more than when a voice with a different personality was used to the user. This proves that an assistant with a similar personality is more likeable and trustworthy. Therefore, personality can be very powerful and an essential tool. It is an inexpensive way to improve the trust and likeness of an assistant by altering the four aspects of voice to suit the user's personality.

	Ms and *SDs*			
	Introvert participant		Extrovert participant	
Dependent variable	Introvert voice	Extrovert voice	Introvert voice	Extrovert voice
Voice extrovertedness	3.41 (1.02)	4.43 (1.73)	3.28 (0.81)	5.30 (1.53)
Liking of the voice	2.82 (1.26)	2.21 (0.92)	1.69 (0.96)	3.08 (1.32)
Credibility of the voice	6.43 (1.46)	5.06 (2.55)	4.18 (1.60)	5.46 (1.18)
Quality of the review	5.44 (0.96)	4.96 (1.31)	4.06 (1.35)	4.79 (1.66)
Liking of the reviewer	5.06 (2.07)	4.74 (2.37)	2.89 (1.27)	5.24 (1.88)
Credibility of the reviewer	4.98 (0.55)	4.52 (0.89)	4.40 (0.70)	5.02 (0.57)
Reviewer extrovertedness	4.68 (1.39)	5.56 (1.37)	4.22 (1.66)	5.77 (1.42)
Buying intention	3.68 (0.98)	3.10 (1.24)	2.93 (1.08)	3.59 (1.17)

Figure 13. Results of the personality experiment. Standard deviations are in parentheses (Nass and Lee 2001).

An important question to ask is how does the assistant learn the user's personalities? Asking questions each time a new user interacts with a voice assistant will take too much time. It indicates that the user does not have a decision in deciding what voice will be chosen. Nass and Brave (2005) give alternative suggestions to better serve the user's needs. One alternative is to ask the user to log in to their Google or Facebook account, which includes personal data. This way, the assistant can analyse the data and predict a suitable personality for the user. However, many users may find this unethical and strange. A more suitable alternative would be to analyse the user's behaviour. As the user speaks to the assistant, it can pick up personality traits while interacting with the user and therefore, it can alter its own personality in real time. The assistant can measure the user's pitch, speed and volume of their voice to match the user's voice and personality. If the device is unable to predict the user's personality, an extroverted voice should be applied (Nass and Brave 2005). This is because users prefer others who are expressive, as they are perceived as friendlier (Friedman et al. 1988). This should however be only applied if the system cannot distinguish the user's personality.

Rather than predicting that a user is an introvert or extrovert, Wiggins (1979) developed a model called the "interpersonal circumplex", to predict if a user is friendly, dominant or vice versa. The friendliness can be determined by the voice's pitch, speed and frequency range, while the dominance of a user can be determined from loudness, the deepness of the voice and

a limited frequency range (Wiggins 1979). Nass and Clifford (2005) also mention that this model is a great benefit as it allows the designer to design more personalities without creating unusual or uncanny voices and personalities. For example, Darth Vader from Star Wars has a slow deep voice which creates a sense of evilness. A henchman has an unfriendly and submissive voice which shows his loyalty but also creepiness. These voices would not work well in a voice assistant.

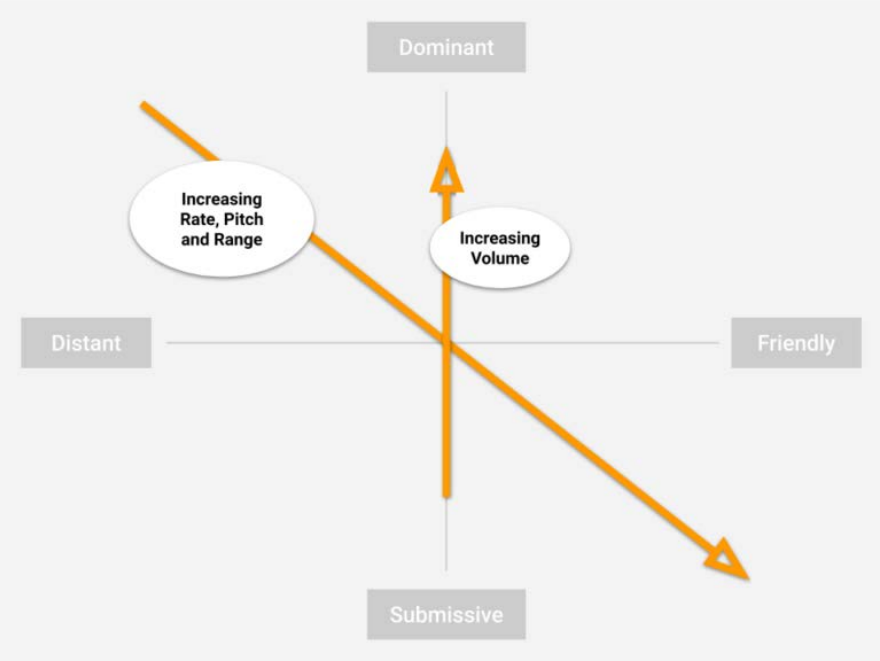


Figure 14. The Wiggins chart plots how voice traits influence personality (Wiggins 1979).

3.6 Is voice cloning the right approach?

Over the last few years, speech synthesis has immensely bridged the gap between human and synthetic speech on a perceptual level. Software exists now that are so realistic to human voice, the human ear is unable to distinguish the difference. Voice cloning is a recent term used to describe a deep-learning algorithm that is able to synthesize a voice to sound very similar to the original inputted voice. The first cloning software emerged in 2016, called Adobe VoCo, a text-based editing tool that allows the user to modify audio narrations by editing or replacing existing words in a text (Jin et al. 2017). Similar applications include Lyrebird AI, CereVoice Me, iSpeech Deepsync and Resemble AI. There are many advances for the use of voice cloning. As mentioned before, it can personalise the user's voice assistants to match their voice, improving the user's experience. It also helps people who have lost their voice who can now gain back a sense of individuality rather than a robotic voice speaking for them. Audiobooks

can also be automatically synthesized, rather than a narrator physically reading the whole book. However, voice cloning can lead to defamation of famous people, as their voice could be manipulated to provoke conflict. People who have access to these types of software could in reality authenticate as another person and access their bank accounts that reply on voice authentication (Vaidya and Sherr 2019). From my analysis, this is the main reason why voice cloning has not been implemented into voice assistants. Before the voice cloning feature can be added, there must be a fully functioning system that can differentiate between a human voice and a synthesis voice in case such an incident was brought to court (AlBadawy et al. 2019).

AlBadawy et al. (2019) are using forensic techniques that can distinguish unusual and specific "spectral correlations not typically found in human speech". Future research needs to be done on finding out what the unusual spectral correlations are, however, the system is currently able to tell the difference between a human and the current speech synthesis application (See figure 14).

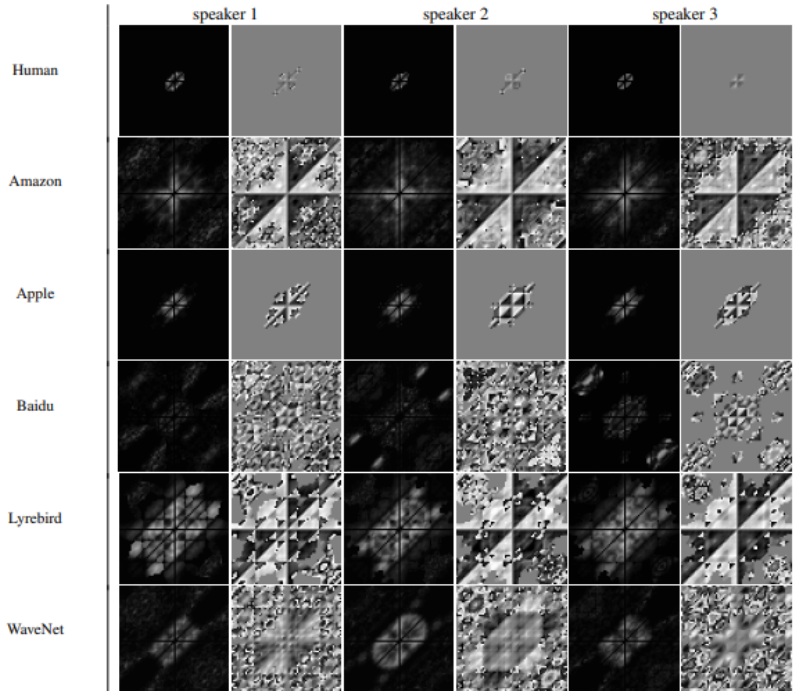


Figure 15. The graph displays a bicoherence magnitude and phase for five synthesized applications and a human speaker. The magnitude and phase are considerably smaller in the human voice compared to the others (AlBadawy et al. 2019).

"VoiceMask" is a tool to aid user's privacy concerns (Qian et al. 2018). The software is an intermediary between the cloud and users that anonymises the user's data before the data is sent to the cloud for speech recognition. Developed on Android, it can resist de-anonymisation

attacks as well as reducing the user's chance of being recognised by a mean of 84%. "VoiceMask" boast that the accuracy of the user's voice is reduced to no more than 14.2% of the original voice (Qian et al. 2018). When the server receives data from the user's voice, it first converts the speech to text and then sends it back before a command is executed. This way the third-party app cannot store and publish the user's data for business and marketing purposes. Not only can it hide user's location, phone information, voice ID etc, it can also protect existing voice inputs from current service providers (Qian et al. 2018).

Figure 16. VoiceMask voice clonability application (Qian et al. 2018).

3.7 Do users want their voice used in a voice assistant?

As explained above, research has shown that user data can be protected when it comes to voice cloning. The better the technology can protect user's, the sooner it can be integrated into voice assistants. However, has the question been asked if the users really want this feature? Voice cloning is a very recent invention, therefore there has not been enough user research to answer this question. Wester et al. (2017) conducted an online quiz to determine what users think is currently possible with speech synthesis technologies. "Bot or Not" was the name of the online tool which could explain what is currently possible and explain how voice modification and speech synthesis have improved recently, using the question "is this recording a bot or not?" (Wester et al. 2017). The online website played a sentence spoken from a famous person (Barack Obama, Stephen Fry, Hilary Clinton) and the user was prompted to click a picture between a robot or the individual. When the user inputs an answer the website displays the correct answer. Out of 144 participants, the average score was 57.6% correct. The highest score was 90% correct while the lowest score was 35% correct. The research did not draw a conclusion; however, it did highlight interesting ethical challenges when implementing realistic voices in voice assistants. The users preferred realistic synthetic voices over robotic

sounding voices, however the users would have rathered if they were instructed beforehand on whether they were listening to a robot or not. This was an interesting point as although most users would distinguish the realistic voice from the assistant, they preferred to know this in advance rather than discovering it themselves.

A major reason for the delay of user research in voice cloning devices is the fact that voice synthesis systems continue to retain a slow conversational flow. As discussed earlier, Wizard-of-Oz experiments are the norm for researching and testing conversational agents. However, this is difficult to do with a voice cloning system, like Lyrebird AI. Voice Puppetry (2020) is a new solution to test users in real time. The system requests one person to speak a sentence, check the accuracy of the phrase and if needed, re-render it. The speed can be changed and saved. When the phrase is rendered, it can be played back on a speaker for an audience to hear (Aylett and Vazquez-Alvarez 2020a). Although the system is still at its development stages, it can interpret what the user is saying in almost real time and play it back in a realistic voice. This is a major step in the process towards realistic conversational voice interfaces.

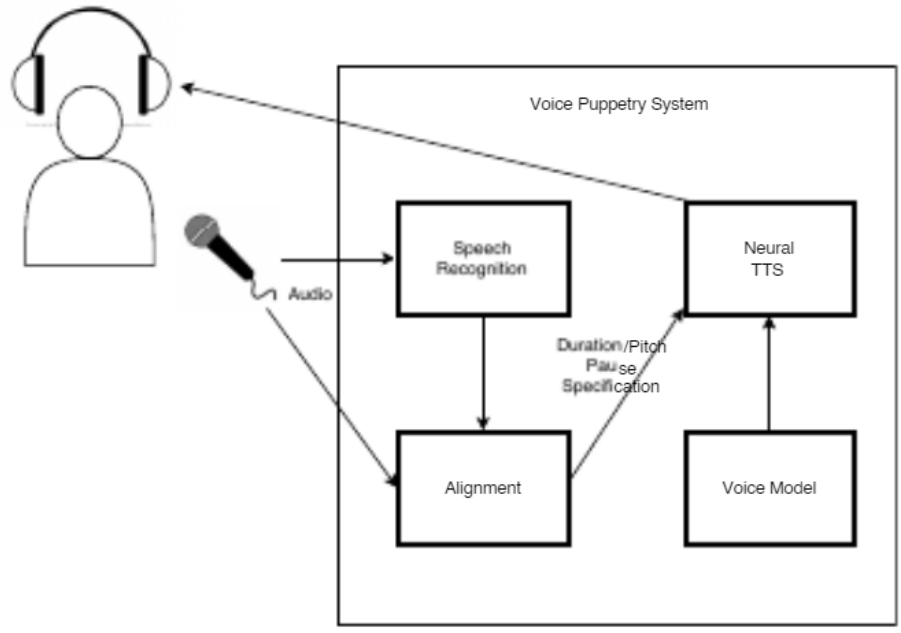


Figure 17. Voice Puppetry speech recognition system (Aylett and Vazquez-Alvarez 2020a)

A recent user study was conducted using the voice puppetry system. A focus group was carried out to discuss voice cloning and how users respond to it. Two members from an engineering and two members from a commercial background were asked to record a phrase and re-record

it until they were satisfied with the result (Aylett and Vazquez-Alvarez 2020b). They were asked what the advantages and disadvantages of this technology might be. One participant called it “having their voice finely controlled”. An advantage a participant pointed out was that the voices sounded more human, which was lacking before. Participants were worried however about faking a person’s identify or for sinister purposes. This experiment concluded that designing a user-centred voice cloning system will be a difficult task. Design issues were raised, such as vocal identity, which can be very sensitive and will require care. The technology is still imperfect, therefore careful design could reduce errors. Finally, allowing the user to modify and review the cloning result is key for user satisfaction (Aylett and Vazquez-Alvarez 2020b).

Figure 18. 4 members of a focus group trying out Voice Puppetry (Aylett and Vazquez-Alvarez 2020b).

Despite very little user centred research in voice cloning, companies have begun to emerge with consumer products with the technology. My Car, My Voice is a product that allows users to record their voices for their in-car voice assistants (Hickman 2019). The company boasts that you can clone your own, a family member or friends voice. They predict a more human-like experience as well as enhancing safety (Hickman 2019). Google has begun to experiment their Google Assistant with WaveNet, while Amazon have begun to use Neural text-to-speech to clone famous actors (Schwartz 2019). Despite the recent attraction from companies to test their products with voice cloning software, there still needs to be more research conducted on understanding users’ thought and reactions to voice cloning technology.

4. Findings/Analysis

Currently, voice assistants all have the same approach of “one voice fits all”. This is obvious in the current popular consumer devices like Amazon Alexa, Google Assistant and Siri, which each have their own unique voice, but share the same voice with every user. This is one of the reasons why users have not trusted or reused these devices. This research has shown that the chosen voice has a crucial impact on the user’s behaviour and use for voice assistants. Research has shown that the chosen voice can have a major influence on the user’s trust and influence on voice assistants (Chiou et al. 2020). Recent research has stated that 41% of users who use voice assistants have trust concerns when interacting with the technology (Olson and Kemery 2019). Certainly, there are many reasons why a substantial number of users have trust issues with this technology, for example, data privacy concerns. However, the same can be applied to emails, text messages and phone calls which may also be tracked without the user knowing. Therefore, why are there more trust issues in voice user interfaces compared to other mediums? The purpose of this research was to emphasize that the voice in conversational agents can have a big impact in improve the user’s experience.

In the past, research has shown that recorded human voices in conversational agents were the user’s preference over the generated synthesized voice. However, the sides have shifted thanks to the recent advances in artificial intelligence and voice cloning. Text to speech synthesis has improved drastically, so much so that it can be hard to distinguish what is or not by listeners. Today, speech synthesis is the preferred option for the voice in conversational agents. Craig and Schroeder (2019) proved this by conducting an experiment on comparing a modern text-to-speech engine, a recorded human voice and an old text-to-speech system from the 2000s. They concluded that modern speech synthesis systems are as effective as human recorded voices. There was not a significant difference in terms of credibility between the recorded and text-to-speech voices, therefore both can be encouraged when designing conversational agents. But of course, the realistic speech synthesized voice will be much preferred for designers and engineers; without the necessity to record a voice, a substantial amount of time and money is saved. Craig and Schroeder’s (2019) research is evidence to prove that speech synthesis should be the way forward, not just in conversational agents, but in audiobooks and online tutorials etc.

4.1 The User-Device-Context Model

Designing the voice for a conversational agent can be a daunting and difficult task. There are an overwhelming amount of implications and concepts that must be considered. Cambre and Kulkarni (2019) have proposed a method for designing smart assistants. The model consists of designing a voice through the following three lenses:

1. User
2. Device
3. Context

Depending on the user's goal, there can be an overlap of all three (See figure 18). They recommend focussing on these three lenses for designing a voice.

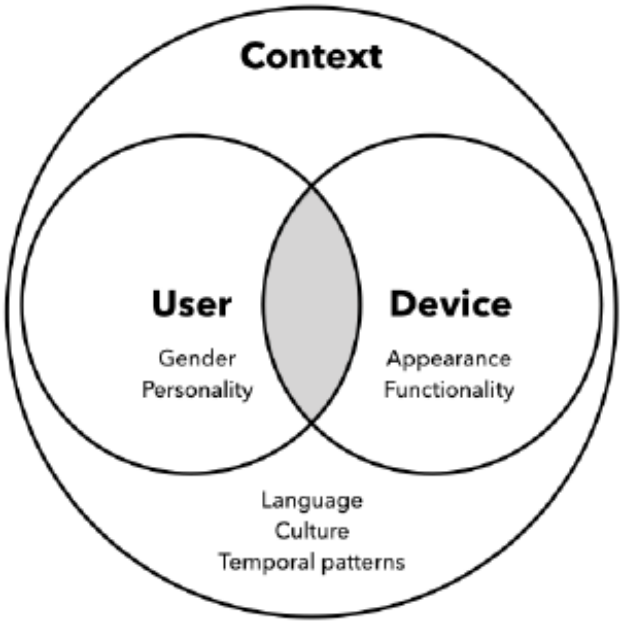


Figure 19. An overview of the model for designing voices for smart devices (Cambre and Kulkarni 2019).

4.2 User

The user was the centre focus for this paper, and how personalising a voice affects the users interaction with a voice assistant. Changing the speech rate, pitch and volume of the synthetic voice to personalise the assistant's voice, making it closer to the user, has shown positive social results in trust, affinity and learning. Lubold et al. (2018) conducted an experiment comparing two robots, one could converse socially and the other spoke without social traits. The result showed that the social robots influenced learning significantly than the non-social. Users were

also more likely to trust an agent that matched the user's personality (Nass and Brave 2005). Research also showed that when designing personality traits in assistants, the user should personally decide on what personalities the agent should have. Users can not only identify personality traits in voice user interfaces, but they are more likely to be attracted to a personality complementary to their own (Lee et al. 2006).

While users prefer agents with similar personality traits, the same can be applied to gender. But how would a genderless voice affect the user's trust in a conversational agent? Research has shown that the majority of users prefer female extroverted voices, when given the option to pick from a range of different voices (Chang et al. 2018). This is the reason why the majority of voice assistants have a default voice of a female extrovert. However, male voices are held more credible by users when talking about information. Users can quickly pick up a voice's gender in only a matter of seconds. Gender has become a concern for designers, as it can have a profound impact on user's interaction with the technology and may lead to social innuendo and stereotypes. The design of voice assistants being labelled as either male or female has reinforced the idea that gender is binary (Cambre and Kulkarni 2019). Genderless voices could be the solution to this ongoing problem. Researchers at Project Q recently built a gender-neutral voice, built from more than twenty participants who identify themselves as non-binary or transgender (Project Q 2019). A survey was conducted to pick a voice that can represent all genders. They suggest this could be an approach for designing voices in assistants in the future, although further research is needed.

4.3 Device

The device itself can have a substantial influence on the user's preference for the voice. The appearance of the device can affect the user's perception of it. Face-face communication is important for learning and processing. Eye gaze, prosody, hand gestures and mouth movements are examples of important cues in communication. These have powerful effects in understanding, supporting speaker thoughts and helping the listener, especially in noisy environments (Grzyb and Vigliocco 2020). Research suggests that users perceive devices with cheeks as feminine and childlike, while devices without a mouth were regarded as unfriendly (Cambre and Kulkarni 2019). Because the voice of a voice assistant is embedded inside a physical device, they suggest the same tendencies apply from the studies of embodied robotics.

For example, the colour of the device might make the user expect the assistant to have a male or female voice. The same can apply to the style, shape, material and size of the device. For example, a small colourful device may be perceived with a child's voice. As mentioned previously, voice systems that sound human-like can mislead users to believe the device can achieve more than it can. Instead of aiming for a more naturalistic voice, designers should match the device with the tasks they are capable of doing. Another interesting aspect to be considered is should the voice in the device should change over time? Just like humans, would changing the devices voice to suit its age improve interactivity? If the device becomes damaged, should the voice also reflect that? Again, this could be something to consider for future research.

4.4 Context

Cambre and Kulkarni (2019) suggest the environment where the user interacts with the voice is the final consideration for designers. This includes linguistic, temporal and cultural factors. For example, Oakley wanted to design a voice for their sport sunglasses. These were no ordinary sunglasses but were designed as a workout tool. The voice needed to represent a workout coach. Because the product was being sold across five different language markets, the designers had to consider cultural and linguistic factors for designing the voice (Danielescu and Christian 2018). Through user testing, it turned out most users preferred an informal tone. However, personality and gender preference differed across the countries and languages. Therefore, they had to design unique voices for each language and culture to match their accents. For example, the French participants would have preferred if the voice started off with a formal voice, but then gradually become more informal the more the user and the computer got to know each other. The Germans on the other hand wanted to get an immediate overview of the computer's credentials in the first interaction (Danielescu and Christian 2018). This shows that there is no global optimal voice that resonates with all users.

To succeed, the designer must consider the "surrounding cultural context" (Cambre and Kulkarni 2019). Linguistic cues are equally as important. Languages can be gendered grammatically, for example, in French, a washing machine is masculine, but feminine in German. Would French speaking people find it strange if their washing machine had a voice of a woman or would German speaking people find it strange having it vies-versa? Siri, Google

Assistant and Alexa all use a similar female voice, therefore would users expect to hear all voice devices to sound the same? As time goes on, this may be an important consideration as voice technology eventually becomes the norm.

Lens	Open Question	Theme
User	To what extent should the voice of a smart device be tailored to its user?	Individualization
User	If a device is used by more than one person, how might it adaptively individualize its voice to match multiple people?	Individualization
User	Under what circumstances should voice characteristics be similar to a user's, and when should voice characteristics be complementary?	Similarity vs. complementarity
User	How does the degree of embodiment of a voice agent affect whether synthesized voices should be similar or complementary?	Similarity vs. complementarity
User	If an individual owns or interacts with several smart devices, should all the devices share the same voice identity, or each speak with a subtly different voice?	Multi-device ecosystems
Device	Could the color, size, stylization, and material form of a smart device trigger stereotypes about the voice's gender or other characteristics?	Physical cues: appearance
Device	Should the voice of a smart device change or adapt depending on its range of motion?	Physical cues: movement
Device	How might the mechanical sounds that smart devices produce (associated with their regular function) shape a user's expectations around a smart device's voice?	Physical cues: sound
Device	How do associations with a device's function (e.g. with traditional gender roles) affect expectations of a device's voice?	Associations with functionality
Device	Should the voice of a smart device change over time?	Long-term use
Device	If a device becomes physically damaged or worn, should that be reflected in the sound of the device's voice?	Long-term use
Context	For users who speak a language with grammatically gender, could the grammatical gender of the noun for a smart device influence the gender that users expect from its voice?	Linguistic cues
Context	As voice assistants like Siri and Alexa grow in popularity, will people come to expect that all voice-enabled devices take on the same, often female-sounding voice?	Longitudinal trends

Figure 20. Future research questions suggested by the User-Device-Context framework for voice devices (Danielescu and Christian 2018).

The User-Device-Context framework is the correct approach designers should take when designing a voice for voice technology. Key questions should be asked beforehand in order to achieve the best possible experience for the user. For example, should the voice be gendered or not, should the device have multiple voices to suit a wider audience and should the voice be robotic or human-like? For the best outcome, the right way forward may be to allow the user to personalise the voice. The current technology, for example, Alexa, Siri and Google Assist, only provide a limited set of voices to choose from. Future devices should allow the user to be in control, to allow the user to decide what gender, personality or age category the voice has, or allow the device to clone the user's own voice. The more choice the user interface can

provide, the more people are satisfied. Only then will we see a much-improved user experience and wider audience.

Conclusion

User centred design is vital for the growth of voice technologies on a consumer level. Research has proven that there are recurring problems with current voice assistants. Cultural norms, social norms, trust and privacy are some of the concerns users have mentioned, when interacting with the current devices. This study has shown that the voice can have a major influence on the user's interaction and thoughts on a voice user interface. If the first interaction with a voice assistant is a bad experience for the user, it can have a long-term impact on the user's interactions with the technology in the future. Current devices all share a common problem in that, one voice fits all. This research has shown that this can have a negative impact on users.

Personality, gender, emotion and cultural norms are some of the traits that need to be considered when designing voice technology. This research has shown that personalities and emotion matching those of the user can improve a user's interaction with the voice. A female voice is the preferred gender however, but this can lead to gender binary. A genderless voice should be considered where possible, to mitigate any potential issues surrounding users feeling excluded. The user's culture and language can also have a big influence on users. The User-Device-Context is a method for designers to consider before designing for voice technology. It focuses on users, devices and contexts to improve the voice. This method will hopefully reduce the current approach of a one voice fits all.

Voice cloning technology was also considered in this research and how it can affect user experiences. While voice cloning can lead to a more personalised and natural experience, it also leads to dangerous and convincing fakes. Voice Mask can be a tool to protect user's data before it is sent to the cloud. If the user's data does get into someone else's hands, bicoherence magnitude, as illustrated, can detect if the cloned voice is real or not. Consumer products using voice technology and user studies have started to emerge, but the technology is still in its infancy. Future research and testing need to be conducted to understand how users will respond to cloning. If not designed correctly, users can find themselves in the Uncanny Valley.

In a perfect scenario, the user should be able to personalise the voice itself or have the voice gradually match the user's traits over time. However, this may not be possible due to cost

issues. An alternative solution would be the use of the user-device-context method in the designing stage.

When voice interfaces drop the current one voice fits all approach, users and computers will begin to speak cooperatively with each other, greatly improving the experience for its users.

References

Al Moubayed, S., Beskow, J., Skantze, G. and Granström, B. (2012) Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive behavioural systems* (pp.114-130). Springer, Berlin, Heidelberg.

AlBadawy, E.A., Lyu, S. and Farid, H. (2019) Detecting AI-synthesized speech using bispectral analysis. *Res. Gate*.

Allen, J., Hunnicutt, M.S., Klatt, D.H., Armstrong, R.C. and Pisoni, D.B. (1987) *From text to speech: The MITalk system*. Cambridge University Press.

Amazon (2017) *Amazon Lex – Build Conversational Bots*, available: https://aws.amazon.com/lex/ [accessed 08 May 2020].

Aylett, M.P. and Vazquez-Alvarez, Y. (2020a) March. Voice Puppetry: Towards Conversational HRI WoZ Experiments with Synthesised Voices. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 69-69.

Aylett, M.P. and Vazquez-Alvarez, Y. (2020b) Voice Puppetry: Speech Synthesis Adventures in Human Centred AI. In Proceedings of the 25th International Conference on Intelligent User Interfaces Companion (pp. 108-109).

Ball, G. and Breese, J. (2000) Emotion and personality in a conversational agent. *Embodied conversational agents*, pp.189-219.

Becker, C., Kopp, S. and Wachsmuth, I. (2007) Why emotions should be integrated into conversational agents. *Conversational informatics: an engineering approach*, pp.49-68.

Bickmore, T. and Cassell, J. (2005) Social dialongue with embodied conversational agents. In Advances in natural multimodal dialogue systems, Springer, Dordrecht pp. 23-54.

Brandtzaeg, P.B. and Følstad, A. (2018) Chatbots: changing user needs and motivations. *Interactions*, *25*(5), pp.38-43.

Burbach, L., Halbach, P., Plettenberg, N., Nakayama, J., Ziefle, M. and Valdez, A.C. (2019) " Hey, Siri"," Ok, Google"," Alexa". Acceptance-Relevant Factors of Virtual Voice-Assistants. In *2019 IEEE International Professional Communication Conference (ProComm)* pp. 101-111.

Cambre, J. and Kulkarni, C. (2019) One Voice Fits All? Social Implications and Research Challenges of Designing Voices for Smart Devices. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), pp.1-19.

Chang, R.C.S., Lu, H.P. and Yang, P. (2018) Stereotypes or golden rules? Exploring likable voice traits of social robots as active aging companions for tech-savvy baby boomers in Taiwan. *Computers in Human Behavior*, *84*, pp.194-210.

Chiou, E.K., Schroeder, N.L. and Craig, S.D. (2020) How we trust, perceive, and learn from virtual humans: The influence of voice quality. *Computers & Education, 146,* p.103756.

Clark, L., Pantidi, N., Cooney, O., Doyle, P., Garaialde, D., Edwards, J., Spillane, B., Gilmartin, E., Murad, C., Munteanu, C. and Wade, V. (2019) What makes a good conversation? challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* pp. 1-12.

Clouse, R.W. and Spurgeon, K.L. (1995) Corporate analysis of humor. *Psychology: A journal of human behavior*.

Cowan, B.R., Pantidi, N., Coyle, D., Morrissey, K., Clarke, P., Al-Shehri, S., Earley, D. and Bandeira, N. (2017) "What can i help you with?" infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* pp. 1-12.

Craig, S.D. and Schroeder, N.L. (2019) Text-to-Speech Software and Learning: Investigating the Relevancy of the Voice Effect. *Journal of Educational Computing Research*, *57*(6), pp.1534-1548.

Csapo, A., Gilmartin, E., Grizou, J., Han, J., Meena, R., Anastasiou, D., Jokinen, K. and Wilcock, G. (2013) Speech, gaze and gesturing: multimodal conversational interaction with NAO robot. In *3rd International Conference on Cognitive Infocommunications,* pp. 667-672.

Danielescu, A. and Christian, G. (2018) A bot is not a polyglot: Designing personalities for multi-lingual conversational agents. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems,* pp. 1-9).

Dudley, H. (1940) The carrier nature of speech. *Bell System Technical Journal*, *19*(4), pp.495-515.

Dudley, H. and Tarnoczy, T.H. (1950) The speaking machine of Wolfgang von Kempelen. *The Journal of the Acoustical Society of America*, *22*(2), pp.151-166.

Edwards, J. and Sanoubari, E. (2019) A need for trust in conversational interface research. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, pp. 1-3.

Ekman, P. (1992) An argument for basic emotions. *Cognition & emotion*, 6(3-4), pp.169-200.

Feng, H., Fawaz, K. and Shin, K.G. (2017) Continuous authentication for voice assistants. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, pp. 343-355.

Fernandes, S., Abreu, J., Almeida, P. and Santos, R. (2018) A Review of Voice User Interfaces for Interactive TV. In *Iberoamerican Conference on Applications and Usability of Interactive TV*, Springer, Cham, pp. 115-128.

Flanagin, A.J. and Metzger, M.J. (2007) The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New media & society*, *9*(2), pp.319-342.

Fogg, B.J. and Nass, C. (1997) Silicon sycophants: the effects of computers that flatter. *International journal of human-computer studies*, *46*(5), pp.551-561.
Fraser, N.M. and Gilbert, G.N. (1991). Simulating speech systems. *Computer Speech & Language*, *5*(1), pp.81-99.

Friedman, H.S., Riggio, R.E. and Casella, D.F. (1988) Nonverbal skill, personal charisma, and initial attraction. *Personality and Social Psychology Bulletin*, *14*(1), pp.203-211.

Grzyb, B. and Vigliocco, G. (2020) Beyond robotic speech: mutual benefits to cognitive psychology and artificial intelligence from the joint study of multimodal communication.

Hickman. (2019) 'Cerence Introduces My Car, My Voice – New Voice Clone Solution to Personalize the In-Car Voice Assistant', *GlobeNewsire*, available: https://www.globenewswire.com/news-release/2019/12/30/1964934/0/en/Cerence-Introduces-My-Car-My-Voice-New-Voice-Clone-Solution-to-Personalize-the-In-Car-Voice-Assistant.html [accessed 08 May 2020].

Hirschberg, J. and Manning, C.D (2015) Advances in natural language processing. *Science*, *349*(6245), pp.261-266.

Holmes, W.J., Holmes, J.N. and Judd, M.W. (1990) Extension of the bandwidth of the JSRU parallel-formant synthesizer for high quality synthesis of male and female speech. In *International Conference on Acoustics, Speech, and Signal Processing*, pp. 313-316. IEEE.

Hoy, M.B. (2018) Alexa, Siri, Cortana, and more: an introduction to voice assistants. *Medical reference services quarterly*, *37*(1), pp.81-88.

Huang, A., Lee, F., Nass, C., Paik, Y. and Swartz, L. (2001) Can voice user interfaces say "I"? An experiment with recorded speech and TTS. *Unpublished Manuscript*.

Jin, Z., Mysore, G.J., Diverdi, S., Lu, J. and Finkelstein, A. (2017) VoCo: text-based insertion and replacement in audio narration. *ACM Transactions on Graphics (TOG)*, *36*(4), pp.1-13.

Juang, B.H. and Rabiner, L.R. (2005) Automatic speech recognition–a brief history of the technology development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, *1*, p.67.

Kepuska, V. and Bohouta, G. (2018) Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home). In *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 99-103).

Khatri, C., Venkatesh, A., Hedayatnia, B., Ram, A., Gabriel, R. and Prasad, R. (2018). Alexa Prize-State of the Art in Conversational AI. *AI Magazine*, *39*(3), pp.40-55.

Kiousis, S. (2001) Public trust or mistrust? Perceptions of media credibility in the information age. *Mass communication & society*, *4*(4), pp.381-403.

Klatt, D.H. (1987) Review of text-to-speech conversion for English. *The Journal of the Acoustical Society of America*, *82*(3), pp.737-793.

Kopp, S., Gesellensetter, L., Krämer, N.C. and Wachsmuth, I. (2005) A conversational agent as museum guide–design and evaluation of a real-world application. In *International workshop on intelligent virtual agents* (p. 329). Springer, Berlin, Heidelberg.

Kostov, V. and Fukuda, S. (2000) Emotion in user interface, voice interaction system. In *Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernetics.'cybernetics evolving to systems, humans, organizations, and their complex interactions'(cat. no. 0* (Vol. 2, pp. 798-803). IEEE.

Lee, K., Peng, W., Jin, S. and Yan, C. (2006) Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction. *Journal of communication*, 56(4), pp.754-772.

Lemmetty, S. (1999) Review of speech synthesis technology. *Helsinki University of Technology*, 320, pp.79-90.

Lison, P. and Kennington, C. (2015) Developing spoken dialogue systems with the opendial toolkit. *SEMDIAL 2015 goDIAL*, p.194.

Lison, P. and Meena, R. (2014) Spoken dialogue systems: the new frontier in human-computer interaction. *XRDS: Crossroads, The ACM Magazine for Students*, *21*(1), pp.46-51.

Lubold, N., Walker, E., Pon-Barry, H. and Ogan, A. (2018) Automated pitch convergence improves learning in a social, teachable robot for middle school mathematics. In *International Conference on Artificial Intelligence in Education*, Springer, Cham, pp.282-296.

Luger, E. and Sellen, A. (2016) "Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* pp.5286-5297.

Lyrebird. (2017) *Ultra-Realistic Voice Cloning and Text to Speech.*

McTear, M., Callejas, Z. and Griol, D. (2016) The Conversational Interface: Talking to Smart Devices: Springer International Publishing, available: *https://doi.org/10.1007/978-3-319-32967-3* [accessed 07 May 2020].

McTear, M.F. (2004) *Spoken dialogue technology: toward the conversational user interface*. Springer Science & Business Media.

McTear, M.F. (2016) The rise of the conversational interface: A new kid on the block? In *International Workshop on Future and Emerging Trends in Language Technology*, Springer, Cham pp. 38-49.

Mead, R. (2017) Semio: Developing a cloud-based platform for multimodal conversational AI in social robotics. In *2017 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 291-292.

Microsoft (2010) *Microsoft Azure*, available: https://azure.microsoft.com/en-us/.
Moon, Y. and Nass, C. (1996) How "real" are computer personalities? Psychological responses to personality types in human-computer interaction. *Communication research*, *23*(6), pp.651-674.

Mori, M. (1970) The uncanny valley. *Energy*, *7*(4), pp.33-35.

Murad, C., Munteanu, C., Cowan, B.R. and Clark, L. (2019) Revolution or Evolution? Speech Interaction and HCI Design Guidelines. *IEEE Pervasive Computing*, *18*(2), pp.33-45.

Nass, C. and Brave, S. (2005) Wired for speech. *How voice activates and advances the human-computer relationship*, Cambridge.

Nass, C. and Lee, K.M. (2001) Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of experimental psychology: applied*, *7*(3), p.171.

Olson, C. and Kemery, K. (2019) *Voice report: Consumer adoption of voice technology and digital assistants*. Technical Report. Microsoft.

Oviatt, S. and Adams, B. (2000) Designing and evaluating conversational interfaces with animated characters. Embodied conversational agents, pp.319-343.

Pearl, C. (2016) *Designing voice user interfaces: principles of conversational experiences.* O'Reilly Media, Inc.

Picard, R.W. (2000) *Affective computing*. MIT press.

Pieraccini, R., Dayanidhi, K., Bloom, J., Dahan, J.G., Phillips, M., Goodman, B.R. and Prasad, K.V. (2003) A multimodal conversational interface for a concept vehicle. In *Eighth European Conference on Speech Communication and Technology*.

Project Q (2019). Meet Q. The First Genderless Voice, available: https://www.genderlessvoice.com/ [accessed 07 May 2020].

Qian, J., Du, H., Hou, J., Chen, L., Jung, T. and Li, X.Y. (2018) Hidebehind: Enjoy Voice Input with Voiceprint Unclonability and Anonymity. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, pp. 82-94.

Reeves, B. and Nass, C.I. (1996) The media equation: How people treat computers, television, and new media like real people and places. Cambridge university press.

Schröder, M., Burkhardt, F. and Krstulovic, S. (2010) Synthesis of emotional speech. *Blueprint for affective computing*, pp.222-231.

Schroeder, M.R. (1993) A brief history of synthetic speech. *Speech Communication*, *13*(1-2), pp.231-237.

Schwartz. (2019) 'New Resemble AI Software Turns 3-Minute Records into Synthetic Speech Profiles', *Voicebot.ai*, available: https://www.globenewswire.com/news-release/2019/12/30/1964934/0/en/Cerence-Introduces-My-Car-My-Voice-New-Voice-Clone-Solution-to-Personalize-the-In-Car-Voice-Assistant.html [accessed 08 May 2020].

Stern, S.E., Mullennix, J.W. and Yaroslavsky, I. (2006) Persuasion and social perception of human vs. synthetic voice across person as source and computer as source conditions. *International Journal of Human-Computer Studies*, *64*(1), pp.43-52.

Vaidya, T. and Sherr, M. (2019) You talk too much: Limiting privacy exposure via voice input. In *2019 IEEE Security and Privacy Workshops (SPW)*, IEEE, pp. 84-91.

VoiceXML (2000) *Introduction*, available: https://voicexml.org/tutorials-2/introduction/ [accessed 07 May 2020].

Wakefield, J, (2016) Would you want to talk to a machine?, *BBC*, available: https://www.bbc.com/news/technology-36225980 [accessed 07 May 2020].

Wang, Y., Skerry-Ryan, R.J., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S. and Le, Q. (2017) Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Weizenbaum, J. (1966) ELIZA---a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), pp.36-45.

Wester, M., Aylett, M.P. and Braude, D.A. (2017) November. Bot or not: exploring the fine line between cyber and human identity. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 506-507.

Wheatstone, C. (1879) The Scientific Papers of Sir Charles Wheatstone. *London:(for Physical Society of London) Taylor & Francis*.

Wiggins, J.S. (1979) A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of personality and social psychology*, *37*(3), p.395.

Zue, V., Seneff, S., Glass, J.R., Polifroni, J., Pao, C., Hazen, T.J. and Hetherington, L. (2000) JUPlTER: a telephone-based conversational interface for weather information. *IEEE Transactions on speech and audio processing*, *8*(1), pp.85-96.

Zue, V., Seneff, S., Polifroni, J., Phillips, M., Pao, C., Goodine, D., Goddeau, D. and Glass, J. (1994) PEGASUS: A spoken dialogue interface for on-line air travel planning. *Speech Communication*, *15*(3-4), pp.331-340.