VRIJE
UNIVERSITEIT
BRUSSEL

Thesis submitted in fulfilment of the requirements for awarding the degree of
Doctor of Engineering Sciences (Doctor in de Ingenieurswetenschappen)

Semantic-Free Affective Speech Framework for Social Human-Robot Interaction

Selma Yilmazyildiz

October 2017

Supervisor: **prof. dr. ir. Werner Verhelst**

Engineering Sciences (Ingenieurswetenschappen)
ETRO – Electronics and Informatics

Examining Committee

Prof. dr. Bart de Boer - Vrije Universiteit Brussel - Chair

Prof. dr. ir. Rik Pintelon - Vrije Universiteit Brussel - Vice-chair

Prof. dr. ir. Hichem Sahli - Vrije Universiteit Brussel - Secretary

Prof. dr. Géza Németh - Budapest University of Technology and Economics - Member

Prof. dr. Khiet Truong - University of Twente - Member

Prof. dr. ir. Werner Verhelst - Vrije Universiteit Brussel - Supervisor

Acknowledgments

I remember my last weeks before flying to Belgium for my masters with my head full of questions about my new life. As a young freshly graduated engineer enthusiastic about science, I was looking for opportunities that I could improve my interest on signal processing during my master studies at VUB. Then I decided to express my interest with an email to the professors at ETRO prior to my move to Belgium... and that is how my journey at ETRO started. Werner Verhelst was one of the first persons immediately reacted upon my interest with a list of possible opportunities. He was kind enough to offer a meeting in my fist days in Belgium to further discuss and so was one of the first persons I've met at VUB. I was impressed by the research topics he could offer and after our first meeting I was more confident that I made a right decision by coming to Belgium for my master studies.

At a second meeting with him, I was looking for a master thesis topic. And once he started talking about a project about emotional vocalization system of a robot for hospitalized children, it touched my hearth and triggered my enthusiasm. Combining my interest in signal processing, the robots that I had only seen in movies so far, and the fascinating opportunity to do science for a good reason: children in hospitals... I couldn't consider any other topic anymore. And my adventure in the world of speech processing and robotics which later also expanded to a PhD started.

Now, standing with a completed thesis which once looked never-achievable, I would like to express my gratitude to the many people who I have met along this journey and who made this PhD possible.

First of all, I would like to thank my supervisor Werner Verhelst who gave me the opportunity to start my academic career with a PhD under his guidance. We spent hours discussing the details of my research, struggles, and experiments. I thank him for all the support, advice and feedback he gave me throughout this entire study.

I would like to express my gratitude to Khiet Truong, Geza Nemeth, Hichem Sahli, Rik Pintelon and Bart De Boer for accepting to be part of my examining committee and contributing with their comments to improving this dissertation. I would also like to thank Roger Vounckx for stepping in promptly when his support

was required.

I also thank my current and past colleagues at ETRO for making it a special place. Lukas, Gio, Henk, Yorgos, Tomas and Wesley thanks for your friendship and the nice working atmosphere. Many of you already started your life out of VUB but I'm sure that we will keep in touch. Special thanks go to my former office mates, Tomas and Wesley for their continued positivity, all the special memories and their contributions to my soft skills including drawing cartoons on a white board, following manly discussions in Dutch and interpreting jokes that were even harder to decode than many of my research questions.

My colleagues at ETRO and beyond, didn't only make this journey a wonderful one but they have also provided specific contributions to this dissertation. In this regard, I would like to thank to:

- Robin Read and Tony Belpaeme for their valuable perspectives and Robin for your collaboration in developing the SFU concept and the related literature review, as well as the cups and balls game code for hybrid vocal communication exploration. Robin, I also appreciate your companionship during my time in beautiful Plymouth. Also thanks to all of the remaining members of the ALIZ-E consortium. Being part of such a large project was a wonderful experience.

- All the members of the HOA16 project team, especially the professors Hichem Sahli, Bram Vanderborgh, Dirk Lefeber, Eric Soetens, Johan Vanderfaeillie and Ann Nowe for their advice and support in all the multidisciplinary aspects related to Probo. I especially would like to acknowledge Bram for his valuable insights and ideas throughout my studies and his contributions in the experiments involving Probo. Thanks to Bram, Jelle, Kristof and David making the experiments with Probo a truly multidisciplinary endeavor across Mechanical Engineering, Cognitive Psychology, Electronics and Informatics departments.

- Special thanks to Hichem Sahli, for partnering with Werner Verhelst in many milestones during my PhD such as introducing me to the ALIZ-E project. Also for his guidance, critical thinking and support throughout my PhD and specifically in designing and executing the experiments performed during the "Robot Week" for children.

- Yorgos Patsis for his expertise and help in performing high-quality recordings and experimental setups as well as for his contributions to the real-time speech modification system for NAO and to the automated database labeling.

- Gio for his committed support in programming and troubleshooting Nao,

Lukas for his deep linguistic expertise, Henk for sharing his broad experience and technical expertise.

I would also like to acknowledge Karin De Bruyn, Luc van Kempen, Irene Raadschelders, Ingrid Sansens, Mike, Pieter, Fengna, Weiyi and many other current and past ETRO members for their support in overcoming challenges throughout my journey.

Many thanks go to my friends (too many to list here but you know who you are!) for providing support and friendship that I needed, also for volunteering to be part of my subjects in many long lasting listening experiments. If Gibberish was a real language, many of you would already be fluent at it.

Lastly, I would like to take this opportunity to thank the most important people in my life, my big family, for all their love and encouragement. I especially thank my mom and dad. They committed their lives for my sister and myself and provided unconditional love and care. Although sometimes hardly understood what I researched on, they were willing to support any decision I made. Thank you for showing such a faith in me. And my sister Berna, being my best friend all my life, and Hakki Abim; I'm grateful to both of you for your trust in me, understanding and encouragement and being always beside me during both the happy and hard moments. I can never forget our never-ending phone conversations with my sister intended for a couple of minutes initially but lasted for hours at the end. I'm also thankful to my parents in-law and sister in-law who have been always supportive and caring since the first moment. Thanks for all your valuable prayers. I know I always have my family to count on when times are rough. I would like to thank my whole big wonderful family who believed in me, who had to get used to the long distance and to being away from us and from their grandson/nephew/cousin Selim.

I'd like to acknowledge a very special person in my family, my husband, Emre for his continued and unfailing love, support, understanding and patience. I could not accomplish this long journey without you by my side. You were always around, at times which were hard and discouraging sometimes and made me go after these periods with your warmth, encouragement and providing new perspectives. I greatly value his contribution that made the completion of this thesis possible and deeply appreciate his belief in me. I also appreciate my little baby boy, Selim, and his amazing patience throughout my lengthy working sessions, especially during my thesis writing. You had to hear "not right now, mommy's working" often – too often. Thanks for tackling with all these and for being such a good boy always making me smile and cheering me up. Thank you, my dear son, thank you for everything that you are, and everything you will become. I look forward to accompanying you

along your own adventure in life.

And finally, there are no words to convey how grateful I'm to God for having all of you around me!

Summary

Recent developments in robotics, artificial intelligence, and machine learning are further accelerating the introduction of robots in our daily lives and the physical environment around us. In order for humans and robots to co-habit in a common space, robots must behave and operate in ways that are similar to or acceptable by humans. In essence, they also need to be social, as we are.

The design, development and study of these social robotic agents and their interactions with humans form up the young but growing field of social Human-Robot Interaction (sHRI). The social robotic agents are equipped with social cues ranging from the use of bodily and facial gestures, natural language and eye gaze to more unique and robot specific methods such as expression through colors, synthetic sounds and vocalizations.

This thesis introduces the umbrella concept of Semantic-Free Utterances (SFU) and brings together multiple sets of studies in social HRI that have never been analyzed jointly before. SFUs are composed of vocalizations and sounds without semantic content or language dependence that may still facilitate rich communication and expression during sHRI. Currently they are most commonly utilized in animation movies (e.g., WALL-E), cartoons (e.g., "Teletubbies,"), and computer games (e.g., The Sims) and hold significant potential for applications in sHRI.

In this thesis, a *Semantic-Free Affective Speech (SFAS) Framework*, which allows robots to express and communicate through vocalizations of meaningless strings of speech sounds (also referred to as affective gibberish speech), has been developed. This framework provides a complete set of tools that can be used as a vocal communication medium for an agent and allows to study diverse aspects of affective human-robot interaction.

As a component of this SFAS framework, a semantic destruction technique that allows a given intelligent text in a certain language to turn into semantic-free gibberish text that is still natural sounding has been developed. Using the methods and techniques outlined, an emotional gibberish speech database (EMOGIB) has been built and made available to the HRI community for further research.

SFAS framework was further enhanced with two modification techniques that are instrumental to utilize the framework across various scenarios in social HRI. One of them is the voice modification capability which provides the alignment of the voice characteristics of the gibberish speech voice with the robot morphology. The second modification, a concatenative synthesis approach which is referred to as segment swapping, decreases the cost of implementation of the framework in HRI studies which will hopefully lead to wider and faster adoption of the framework by the HRI community.

Piloting the implementation of the outlined Semantic-Free Affective Speech (SFAS) Framework, sets of experiments that assess the effectiveness of using SFAS across various aspects of affective human-robot interaction were performed. The results of these experiments have shown the expansive applicability of the proposed framework in social HRI, while outlining certain improvement areas in various components used in the pilot implementations.

Publications

Journal papers

- **S. Yilmazyildiz**, R. Read, T. Belpeame and W. Verhelst, "Review of Semantic-Free Utterances in Social Human-Robot Interaction", *International Journal of Human-Computer Interaction*, Vol: 32(1), pp: 63-85, ISBN-ISSN: 1044-7318, 2016.

- **S. Yilmazyildiz**, W. Verhelst and H. Sahli, "Gibberish speech as a tool for the study of affective expressiveness for robotic agents", *Multimedia Tools and Applications*, Vol: 74(22), pp: 9959–9982, ISBN-ISSN: 1573-7721, 2015.

- **S. Yilmazyildiz**, D. Henderickx, B. Vanderborght, W. Verhelst, E. Soetens and D. Lefeber, "EMOGIB: Emotional Gibberish Speech Database for Affective Human-Robot Interaction", *Lecture Notes in Computer Science*, Vol: 6975, pp: 163 - 172, ISBN-ISSN: 978-3-642-24570-1, 2011.

- **S. Yilmazyildiz**, L. Latacz, W. Mattheyses and W. Verhelst, "Expressive Gibberish Speech Synthesis for Affective Human-Computer Interaction", *Lecture Notes in Artificial Intelligence*, pp: 584 - 590, ISBN-ISSN: 978-3-642-15759-2, 2010.

- J. Saldien, K. Goris, **S. Yilmazyildiz**, W. Verhelst and D. Lefeber, "On the Design of the Huggable Robot Probo", *Journal of Physical Agents*, Vol: 2(2), pp: 3-12, ISBN-ISSN: 1888-0258, 2008.

- **S. Yilmazyildiz**, W. Mattheyses, G. Patsis and W. Verhelst, "Expressive Speech Recognition and Synthesis as Enabling Technologies for Affective Robot-Child Communication", *Springer Lecture Notes in Computer Science*, Vol: 4261, pp: 1-8, ISBN-ISSN: 0302-9743, 2006.

Conference papers

- W. Wang, G. Athanasopoulos, **S. Yilmazyildiz**, G. Patsis, V. Enescu, H. Sahli, W. Verhelst, A. Hiolle, M. Lewis and L. Canamero, "Natural Emotion

Elicitation for Emotion Modeling in Child-Robot Interactions", *Proceedings of the 4th Workshop on Child Computer Interaction (WOCCI 2014)*, 2014.

- **S. Yilmazyildiz**, D. Henderickx, B. Vanderborght, W. Verhelst, E. Soetens and D. Lefeber, "Multi-Modal Emotion Expression for Affective Human-Robot Interaction", *Workshop on Affective Social Speech Signals (WASSS 2013)*, 2013.
- **S. Yilmazyildiz**, G. Athanasopoulos, G. Patsis, W. Wang, M.C. Oveneke, L. Latacz, W. Verhelst, H. Sahli, D. Henderickx, B. Vanderborght, E. Soetens and D. Lefeber, "Voice Modification for Wizard-of-Oz Experiments in Robot-Child Interaction", *Workshop on Affective Social Speech Signals (WASSS 2013)*, 2013.

Extended abstracts

- **S. Yilmazyildiz**, G. Patsis, W. Verhelst, D. Henderickx, E. Soetens, G. Athanasopoulos, H. Sahli, B. Vanderborght and D. Lefeber, "Voice Style Study for Human-Friendly Robots: Influence of the Physical Appearance", *5th International Workshop on Human-Friendly Robotics (HFR2012)*, 2012.
- G. Athanasopoulos, W. Wang, F. Wang, **S. Yilmazyildiz**, M.C. Oveneke, V. Enescu, H. Sahli and W. Verhelst, "Towards Autonomous Child-Robot Interaction", *5th International Workshop on Human-Friendly Robotics (HFR2012)*, 2012.

Edited book chapter

- K. Goris, **S. Yilmazyildiz**, J. Saldien, B. Verrelst, W. Verhelst and D. Lefeber, "Probo, a friend for life?", *Brave New Interfaces: Individual, Social and Economic Impact of the Next Generation Interfaces*, ISBN-ISSN: 978-90-5487-416-4, 2007.

Contents

List of Figures

List of Tables

Abbreviations

ALIZ-E	Adaptive Strategies For Sustainable Long-Term Social Interaction
AMDF	Average Magnitude Difference
CRI	Child-Robot Interaction
DOF	Degrees Of Freedom
DSP	Digital Signal Processing
EMOGIB	Emotional Gibberish Speech Database
GS	Gibberish Speech
HRI	Human-Robot Interaction
IDE	Integrated Development Environment
LED	Light Emitting Diodes
LPC	Linear Predictive Coding
MLEIR	Musical Language For Emotional Interaction Between Robots
MOS	Mean Opinion Scores
MU	Musical Utterances
NLI	Natural Language Interfaces
NLP	Natural Language Processing
NLU	Non-Linguistic Utterances
PLAR	Pseudo Log Area Ratio
PU	Paralinguistic Utterances
RMS	Root Mean Square
ROILA	Robot Interaction Language
SAM	Self-Assessment Manikin
SDK	Software Development Kit
SFAS	Semantic Free Affective Speech
SFU	Semantic Free Utterances
sHRI	Social Human-Robot Interaction
STFT	Short-Time Fourier Transform
TTS	Text-To-Speech
WoZ	Wizard-Of-Oz
WSOLA	Waveform Similarity Based Overlap-Add

1 Introduction

1.1 Motivation

Since the introduction of the Unimate robot into the General Motors assembly line in the 1950's, robots have gradually undergone numerous changes with respect to their physical design as well as their utility across various application domains. In short, the life of a robot is no longer confined to the enclosed work cells of an assembly line. Rather, robots can be found operating in a variety of environments and applications using different degrees of autonomy as well as coming in all shapes and sizes. For example, swarms of robots can be found racing on the floors of warehouses, collecting and analyzing soil samples on distant planets, examining shipwrecks in the depths of our oceans, and more recently, driving us between different locations. The general emerging trend is that robotic technology is slowly, but surely, making its way into our daily lives, and beginning to share the same physical space with us and as a result coming into direct contact with the general population.

This sharing of the same physical spaces has important implications regarding the design of robots with respect to their mechanical construction and their programming. Concretely, robots need to be designed to cater for safe interactions with people and they need to understand and be sensitive to how people behave. Similarly, they need to be designed such that people understand what the internal state of the robot is through its observable behaviour. Essentially, in order for humans and robots to co-inhabit a common space, robots must behave and operate in ways that are similar to humans. In essence, the argument is that they too need to be social, as we are (Breazeal, 2002).

To be able to address the need to be social, these robotic agents should be empowered with various affective social cues enabling natural and intuitive social interactions with humans. While from these social cues the first ones that come to mind might be bodily and facial gestures or natural language, which are also affective social cues humans utilize, there are other applicable methods which are more unique to robots, like expression through colors, synthetic sounds and vocalizations (Breazeal, 2002; Embgen et al., 2012). Such sounds and vocalizations may not necessarily involve semantics in natural spoken language and are referred

to as Semantic-Free Utterances (SFU) (Yilmazyildiz, Read, Belpeame, & Verhelst, 2016).

The design, development and study of these social robotic agents and their interactions with humans form up the young but growing field of social Human-Robot Interaction (sHRI). Like many technologies that were considered futuristic at some point in time, entertainment industry and specifically science-fiction and animation movies didn't only imagine and explore how the future could look like together with these social robots, but also created some expectations and in some cases even stereotypes about social robots of the future. For example, R2D2 from the Star Wars movies, and the robots WALL-E and Eve from the Disney-Pixar movie WALL-E, show us that robots do not need to leverage natural language in order to be able to communicate effectively and they can use SFUs instead. These robots employ a rich repertoire of beeps, squeaks, whirrs and nonsense vocal utterances, which are SFUs, to engage in stimulating and entertaining social interactions with great success, and more importantly, the audience is unfazed by the use of these alternative methods of communication. Clearly there is inspiration that may be taken from the world of animation and film and how they have brought their robotic characters to life.

The thesis presented here is concerned with one of such methods that allow robots to express and communicate through vocalizations of meaningless strings of speech sounds that may still facilitate rich communication and expression during HRI - namely Gibberish Speech.

Natural Language Interfaces (NLI) have long been an important horizon goal of Human-Machine Interfaces and the technologies behind such interfaces (namely Speech Recognition, Natural Language Understanding, Natural Language Generation and Speech Synthesis) have been subject to much research, design and development over the last few decades (Theobalt et al., 2002; Imai, Hiraki, Miyasato, Nakatsu, & Anzai, 2003; Jung et al., 2005; D'Mello, McCauley, & Markham, 2005; Mozos, Jensfelt, Zender, Kruijff, & Burgard, 2007; Gorostiza & Salichs, 2011; Connell, 2014). However, the current state of the art in these technologies is still far from the spoken language capabilities of an average human speaker or listener, especially in adverse real-world situations (Moore, 2014). Herein lies an important problem for the uptake of socially capable robotic systems in the near future. The rate of development and deployment of social robotic systems that may interact with people is so rapid that the rate of advances in Natural Language Interfaces may not be able to catch up. The result of this is that, currently, state of the art social robots are unable to leverage the full power of natural language.

For dealing with situations where NLI might fail during an interaction, a number of strategies have been explored: constraining and even scripting interactions and dialogues or narrowing the scope of responses expected from the user (e.g., Lohse, Rohlfing, Wrede, & Sagerer, 2008); requesting the user to repeat the input if the recognition result is inconsistent with the dialog discourse (e.g., Holzapfel & Gieselmann, 2004); asking clarification questions (e.g., Gabsdil, 2003; Deits et al., 2013); or even employing a set of general purpose responses to try and catch the failing interaction (e.g., Lison & Kruiff, 2009). Such strategies have the risk of revealing the limitations of the system to the users (Ros Espinoza et al., 2011) as unrelated, incorrect or repetitive answers are easily spotted by the users. That is one of the roadblocks in front of the development of long-term, open-ended HRI which is the outstanding goal of the field (Belpaeme et al., 2012). As such camouflaging these limitations from the users by reverting to a replacement modality so that the interaction can continue, even if with some limitations, could be appealing.

Not having any semantic information implies obvious limitations for SFUs when compared to Natural Language. However, they also have important qualities that make their implementations quite promising in sHRI. As an example, SFUs are not bound to a specific accent of a specific language. Thus the use of SFUs in multi-lingual and/or multi-cultural environments is advantageous where dealing with speakers with foreign accents is currently very challenging for NLI (Moore, 2014).

Also the SFUs are less demanding for computing power than NLI systems which helps addressing the system response time requirements that are important for human-like sHRI (Shiwa, Kanda, Imai, Ishiguro, & Hagita, 2009). Moreover, SFUs having less content to be decoded by the user, the interpretation in combination with the context of the interaction and situation is mostly left to the users. Especially children do not see robots as mechanical machines, and they readily anthropomorphise robots and maintain the illusion that they have life-like characteristics (Belpaeme et al., 2013) which creates another advantage for the use of SFUs.

Driven with all these motivations, the objective of this thesis is to provide a framework that allows to study affective human-robot interactions by using vocalizations that do not involve semantics in natural spoken language, so-called Affective Gibberish Speech or Semantic-Free Affective Speech (SFAS). The high-level architecture of the framework is illustrated in Figure 1.1. First, the strategy of creating this framework will be described along with the evaluations on isolated audio utterances in the first chapters. That will then be followed by utilization of the framework in pilot implementations to real world social robots and providing insights on the potential usage while seeking answers to questions related to its

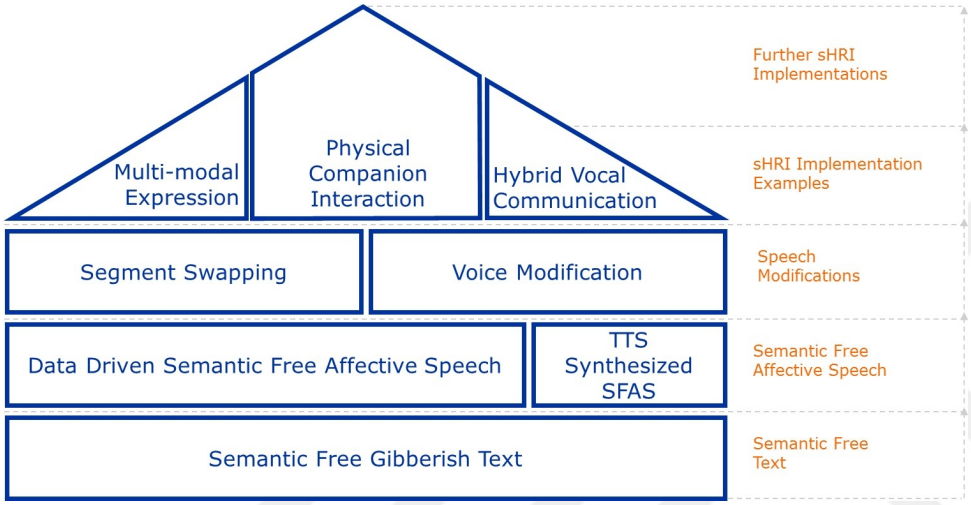


Figure 1.1: Semantic-Free Affective Speech Framework Architecture

deployment in the final chapters.

1.2 Contributions

As a result of the research presented in this thesis, a number of important findings have been uncovered, which are listed below as the key scientific and technical contributions made by this work:

- **The concept of Semantic-Free Utterances (SFUs) has been introduced:** Sounds and vocalizations that do not involve semantics in natural spoken language such as Gibberish Speech, Non-Linguistic Utterances, Musical Utterances and Paralinguistic Utterances are brought together under the umbrella-term, Semantic-Free Utterances. By introducing the concept of SFUs and bringing together multiple sets of studies in social HRI that have never been analyzed jointly before, the need for a comprehensive study of the existing literature for SFUs is addressed, the current grand challenges and open questions are outlined and guidelines are provided for future researchers.

- **A Semantic-Free Affective Speech (SFAS) Framework has been developed:** This framework provides a complete set of tools that can be used as a vocal communication medium for an agent that then allows to study diverse aspects of affective human-robot interaction.

- **Development of a semantic destruction technique that allows a**

given intelligent text in a certain language to turn into unintelligible/semantic free text that is still natural sounding: This mechanism, as a component of the SFAS framework, is based on an intelligent swapping strategy that replaces vowel nuclei and consonant clusters of a given text in a language in accordance with the natural probability distribution of the vowel nuclei and consonant clusters of the original language. This results in natural sounding gibberish and still resembles the source language when a good quality synthesis is used.

- **An emotional gibberish speech database (EMOGIB) has been built and made available to the HRI community for further research:** With the methods and techniques in the SFAS framework an expressive database that is in gibberish form and sounds like a real language was created. The resulting EMOGIB database was also utilized as a component of the SFAS framework in its pilot implementations. Contributions with the database building also include configuring a set-up strategy for maintaining constant recording conditions and steady voice type/quality throughout each emotion category.

- **Showcased the direct relation between the physical appearance of the robots and the appropriate voice pitch:** The lower pitched voices are perceived more related with the high volume (i.e. larger) robots while the higher pitched voices are found to be more related with the low volume (i.e. smaller) robots. The feature of voice modification is included in the SFAS framework, which provides the ability to easily perform the required voice alignment.

- **Improved resolution of the ambiguities and confusions in facial expressions of a robotic agent by the presentation of semantic-free gibberish speech:** The emotional information exchange with robots takes place in different layers of multimodal interaction. However, it was not known whether the effect of the speech without semantic meaning on the emotion expression would be positive or negative. A multi-modal evaluation study showed that gibberish speech improves the emotion recognition significantly.

- **Semantic-Free Affective Speech can be used as the sole vocal medium or in combination with a natural language in affective HRI implementations:** In the preference rankings between only natural language, only gibberish and mixtures of natural language and gibberish, no statistically significant differences were found. This result implies that Gibberish can be used as the sole vocal medium or in combination with a natural language in sHRI studies.

- **Development of a concatenative synthesis approach that further en-**

hances the capabilities of the framework with no significant negative effect on emotion recognition and acceptable levels of drop in naturalness: By swapping the units of an utterance with other units from the database of the related emotion, this synthesis capability expands the number of unique semantic-free utterances.

1.3 Outline

This section outlines the structure of this dissertation along with a brief description of the theme and context for each chapter.

Chapter 2 introduces and defines the concept of Semantic-Free Utterances, under which semantic-free gibberish speech is categorized, and provides a comprehensive literature overview of the field. Finer details of each of the four SFU categories (Gibberish Speech, Non-Linguistic Utterances, Musical Utterances and Paralinguistic Utterances) are given and the developments that have been made in the state-of-the-art, as applied to HRI, are charted, while at the same time positioning this thesis in the literature. This is followed by a summary and discussion that highlights the areas of success and sketches general areas that require more research as well as the current grand challenges, open questions and future directions, that this area of research faces.

Chapter 3 describes the approach used in removing the semantic content in a given text, thus creating semantic-free gibberish text. The chapter concludes by representing two experiments evaluating the perceived naturalness and emotion conveying capabilities of the resulting gibberish.

In *chapter 4*, the design and building procedure of the emotional gibberish speech database (EMOGIB) that is used in the data-driven method of semantic-free gibberish speech synthesis is explained. The chapter also presents the assessment of the quality of the emotions as well as the naturalness of the utterances.

Chapter 5 elaborates on two modification techniques that are instrumental to further utilize the Semantic-Free Affective Speech (SFAS) framework in social HRI. One of them is the voice modification algorithm which provides the alignment of the voice characteristics of the gibberish speech voice with the robot morphology and with the voice of the robot's build-in text-to-speech synthesizer. The other modification, a concatenative synthesis approach which is referred to as segment swapping, decreases the cost of implementation of the framework in HRI studies which will hopefully lead to wider and faster adoption of the framework by the HRI community.

In *chapter 6*, pilot implementations focusing on the usage of the SFAS framework with the intended robotic embodiments are presented. It is investigated whether the effect of the speech without semantic meaning on the emotion expression would be positive or negative in multi-modal presentations. The potential use of semantic-free gibberish speech alongside natural spoken language is also assessed as a hybrid vocal communication strategy and, finally, the utilization of gibberish speech is investigated in real-life interaction scenarios with a co-viewing companion scenario by using the embodied robot.

Chapter 7 provides a summarizing overview of the work that has been presented, and concludes the thesis by discussing the results obtained and elaborating on a collection of topics that are potentially fruitful future research.

2 Auditory Affect Synthesis in HRI

2.1 Introduction

The content of this chapter is based on our publication (Yilmazyildiz et al., 2016).

The chapter provides a comprehensive literature overview of the methods that allow robots to express and communicate through sounds and vocalizations (which are referred to as *utterances* in this thesis) that do not involve semantics in natural spoken language but may still facilitate rich communication and expression during HRI. Such utterances can come in four general flavours: Gibberish Speech (GS), Non-Linguistic Utterances (NLUs), Musical Utterances (MU) and Paralinguistic Utterances (PU), all of which are brought together under the umbrella-term *Semantic-Free Utterances (SFUs)*.

Research into SFUs, as applied to social HRI, has received very little attention in comparison to the other modalities of expression (namely affective speech synthesis, facial gestures, gaze cues, and body language). The research efforts that have been undertaken have been varied and scattered. Addressing this, in this chapter, the past developments are charted and a review of the state of the art in these Semantic-Free Utterances is provided.

2.2 Semantic-Free Utterances

2.2.1 Description and subject area

In broad terms, Semantic-Free Utterances (SFUs) can be described as auditory communication or interaction means for machines that allow the expression of emotion and intend, composed of vocalizations and sounds without linguistic semantic content.

Figure 2.1 illustrates the realm of the SFU concept as is dealt with in this

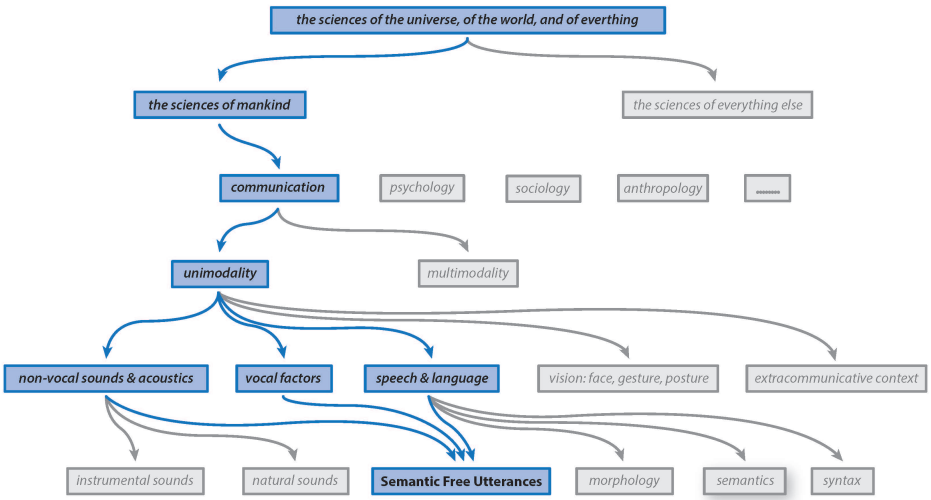


Figure 2.1: Schematic illustration of Semantic-Free Utterances (adapted from (Schuller & Batliner, 2014))

thesis. In the figure, the "science of everything" is narrowed down to the science of mankind, communication and unimodality, as depicted by Schuller and Batliner (2014). Then the notion of SFU is introduced as a composition of auditory attributes of uni-modal communication, namely speech and language, vocal factors and non-vocal sounds and acoustics.

In this graph, the term "language" refers to "natural language", which is modelled and processed within computational linguistics, and speech refers to "spoken language" that is the object of speech processing technology (Schuller & Batliner, 2014). Speech and language research, in this context, deals with the various branches of phonetics and linguistics, such as syntax, semantics, etc.

The "vocal factors" consist of various aspects regarding the human voice. For example, organic vocal aspects such as the difference in the size of the speech organs that affects the pitch of the voice that then characterizes the male/female or child/adult voice. Other vocal factors such as loudness, rate, pitch contour, voice quality contribute to expressive aspects of the human voice.

The field of acoustics deals with topics such as vibration, sound, ultrasounds and infrasound. However, in this graph the focus area of acoustics is restricted to non-vocal sounds in the audible frequency range, which include instrumental and natural sounds as well as computer generated and re-created sound waves and

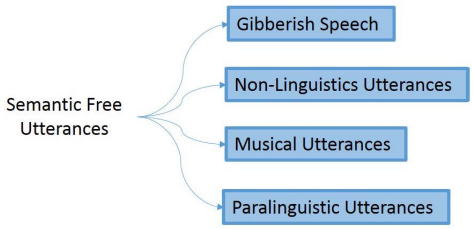


Figure 2.2: Categorization of Semantic-Free Utterances

sound effects.

Having described the main terminologies used in the graph, SFUs can easily be defined *ex negativo*: SFUs comprise everything that is not the focus of linguistical semantics, syntax and morphology in natural language and that does not include instrumental and natural sounds in acoustics and music.

Depending on the differences in the underlying nature and the usage in HRI studies, SFUs can be categorized under two general types: Gibberish Speech (GS) and Non-Linguistic Utterances (NLU). Apart from these, there are other SFU types such as Musical Utterances (MU) and Paralinguistic Utterances (PU) (Figure 2.2). Although paralinguistic research has been receiving more attention recently in the speech processing domain, its utilization in HRI studies has been very limited. Musical Utterances on the other hand, have been employed in more studies than PU, and stand as being one of the sources of inspiration to NLU research.

There are commonalities between GS and PUs as well as between NLUs and MUs. GS and PUs are both utterances that resemble vocalizations of human speech. GS consists of vocalizations of meaningless strings of speech sounds and thus resembles the timbre and voice quality of human speech, without containing any semantic content. PUs on the other hand are non-speech vocal events and thus contain any type of vocal sounds beyond speech (such as laughs, sighs, etc.). In contrast, NLUs and MUs are both non-vocal recreated sound waves/sound effects. NLUs consist of beeps, squeaks and whirrs, and are the sort of auditory signals that do not resemble natural language and speech. MUs sound very similar, and are often synthesized using the same tools as for NLUs. Where the two approaches differ is in

the underlying theories used to drive their creation. MUs use musical theory as the sole source for defining the acoustic properties and dynamics of utterances, whereas NLUs are far less specific in this regard. All these different SFU categories will be elaborated in Section 2.3 along with their underlying theories and techniques.

2.2.2 Utility as a tool in broader HRI

There are already a number of robotic systems, both fictional and non-fictional, that demonstrate how a variety of the qualities of SFUs can be applied to social robots. However, there are also examples that stem from social HRI research showing that SFUs are not only a useful means of facilitating expressive displays, but they may also be used in a manner that helps advance research in other areas of HRI.

Kismet (Breazeal, 2002) is a prominent example of GS being used as a means of vocal expression in a robot, while also serving as a tool to help facilitate research into other areas of HRI simultaneously. For example, evaluating the influence that affective models of the robot's internal states can have on the observable behaviour of the robot, both visual and vocal.

There have been also efforts to utilize GS in robotic agents in the form of language games to understand the evolution of language and language acquisition (Steels, Kaplan, McIntyre, & Van Looveren, 2002; Steels, 2003). The idea was to explore how a population of robotic agents would be able to develop their own communication strategies and their own vocabularies by tackling real-world problems (such as background noise, shifts in the meanings, new lexicons/words etc.) in the real environments, all without human intervention or prior specification. Again related to languages, Mubin et al. (Mubin, Bartneck, & Feijs, 2009) investigated the feasibility of creating artificial languages, specifically designed to optimize the performance of Speech Recognition. With the artificial language they developed (Mubin, Bartneck, & Feijs, 2010), the Robot Interaction Language (ROILA), they aimed at reaching a high level of speech recognition on the robot side and a minimal effort needed in learning this new language on the user side.

In research focused upon the physical, anthropomorphic design of robots and how this impacts the perception people have of a robot, Walters et al. (Walters, Syrdal, Dautenhahn, te Boekhorst, & Koay, 2007) used NLUs to facilitate vocal animation of a robot that was deemed to be "machine-like", as opposed to having a more anthropomorphic design. In this example, NLUs have been used as a tool in order to be able to study aspects of HRI that fall far beyond only affective expression via sound. Another example of the use of NLUs in research is with the robot Keepon (Kozima, Michalowski, & Nakagawa, 2009), which is a robotic tool

designed to be used with young autistic children, many of which are pre-verbal. Here, as the use of natural language with pre-verbal infants serves little purpose and as the morphology is very abstract, NLUs provide an adequate compromising solution. Similarly, Vazquez et al. (Vazquez, Steinfeld, Hudson, & Forlizzi, 2014) have used NLUs to animate a robot "side-kick", Blink, while studying the effects of having a robot side-kick during child-robot interaction. It was a design decision to have a side-kick that did not utilise real natural language as this made the overall technical design of the robotic systems easier to implement as well as falling more in line with the morphology of the robot (which was a lamp shade in this case).

In these examples, the benefit of SFUs is clear. The use of natural language in robots is currently cumbersome due to the limitations that the technology has, and if the research does not strictly require natural language, but does require some form of vocal expression or interaction, SFUs can be seen as an attractive option that are easy to implement compared with NLP.

2.3 Affective interaction with SFUs in social HRI: an overview

Having outlined the motivations and benefits of using SFUs in HRI, in this section the attention is on the finer details of SFUs. Social robots should communicate and interact with humans and may become a companion for humans. As described by Libin and Libin (2004), emotion is one of the major communication layers that make an artificial partner a good companion for humans. This emotional communication and interaction requires recognition, interpretation, processing, and simulation of human affects by the robot. Depending on the shape and the functionalities of the artificial partner, the affective communication can be carried out by means of visual and auditory coding, which then needs to be decoded by the human users.

Visual coding includes encoding the affective information into facial expressions, body gestures or colours of the robot. There are various differences amongst these modalities. For example Kismet (Breazeal, 2002) uses a fully actuated head with 18 degrees of freedom (DOF) to express facial expressions, while eMuu (Bartneck & Michio, 2001) uses far fewer DOF. QRIO (Brooks & Arkin, 2007) uses body gestures while Nao supports its body gestures by also giving meaning to the colours (Haring, Bee, & André, 2011). The pictures of these robots with different affordances can be seen in Figure 2.3.

In the auditory channel, speech with natural language is the primary communication strategy utilized for HRI. What is said and how it is said are dual encoded

Figure 2.3: Examples of robots with different affordances for expressive communication

and transmitted via the same channel in the human voice with speech (Picard, 1997; K. Scherer, 1986, 2003). This makes the study of affective communication via the human voice particularly challenging as disentangling what has been said from how it has been said is not an easy task.

In theory affective expression in the human voice and through music share a common origin from an evolutionary perspective (K. Scherer, 1995), mainly with respect to the use of the voice (i.e. singing), but this can also be extended to the use of musical instruments (Juslin & Laukka, 2003). As such, it makes sense to touch upon the expression of affect through music also. Just like expressive human speech, musical expression has also been explored for a number of years and as a result a large body of research has also accumulated. Juslin and Laukka (2003) found that there are indeed a great number of similarities between the acoustic cues in music and in the human voice when it comes to conveying a particular affective state. For example, when conveying anger, characteristics of a musical piece are found to have a fast rate, have a high intensity with a great deal of variability in this intensity, a high overall pitch with a high variability and fast onsets of notes. Similar characteristics are also found for the expression of happiness, while sadness was associated with a slow overall tempo, a lower pitch with less variability and less overall intensity in the acoustic signal and with less aggressive onset of notes. This is generally consistent with the findings in the human voice. With regard to the notion of whether the human voice and music share a common origin with respect to emotional expression, Juslin and Laukka (2003) conclude that expression of affect through music is likely based around the manner through which it is done in the human voice. As such, the potential use of insights gained from both fields can be considered for the application of creating SFUs.

As stated earlier, SFUs are categorized under four general types (see Figure 2.2): Gibberish Speech, Non-Linguistic Utterances, Musical Utterances and Par-

alinguistic Utterances. In this section, each of the four SFU types will be described, together with their roots in other fields of science, such as Psychology, Musicology, etc. Following this, the main strategies that may be adopted in order to produce affect-laden utterances for social robots will be outlined. However, only the works that are directly relevant to the use of SFUs as applied to the field of social robotics is reviewed. That is, how SFUs can be utilised by robots as social cues in order to communicate information to human users.

2.3.1 Gibberish speech

The term gibberish stands for "Unintelligible or meaningless speech or writing"[1]. In practical terms, it may stand for a structured encrypted language which may seem nonsensical to outsiders (Gardner, 1984), or for vocalizations that sound like speech but have no actual meaning, by masking (K. Scherer, Koivumaki, & Rosenthal, 1972; K. Scherer, 1985; Remez, Rubin, Pisoni, & Carrell, 1981) or manipulating (Cahn, 1990; Burkhardt & Sendlmeier, 2000) the linguistic cues (K. Scherer, 2003)), or for speech or writing that is grammatically, syntactically and phonetically correct but semantically irrational (Chomsky, 1956).

In essence, there is no semantic content in gibberish vocalizations. Several approaches exist to generate gibberish without the semantic content. Some of them use text transcriptions and others operate on the speech signal (a summary of this categorization can be seen in Table 2.1). An early example of the former is from Chomsky (1956). He destroyed the semantic content at the sentence level by utilizing words in a nonsensical combination. The well-known Chomsky sentence "Colorless green ideas sleep furiously" is grammatically correct, but it doesn't convey an understandable meaning. Although his intention was to demonstrate the difference between syntax and semantics, such semantically-anomalous pseudo utterances were then used in emotion decoding, vocal affect measurement studies (K. Scherer, 1986; Pell, Paulmann, Dara, Alasseri, & Kotz, 2009; Paulmann & Pell, 2011) and in testing speech intelligibility (Arslan & Talkin, 1998).

In Chomsky's approach the individual words are still intelligible, but the overall meaning becomes nonsensical once they are specially formed up into phrases with selected words. Thus the language is still recognizable. In other approaches that operate on the text, the semantic is destroyed already at the word level: Jabberwocky sentences (Silva-Pereyra, Conboy, Klarman, & Kuhl, 2007) and nonword phrases (Rastle, Harrington, & Coltheart, 2002) are a few examples which are mainly driven by neurolinguistic interests while mono/poly-alphabetic substitution, playfair cipher, and transposition cipher techniques are used in cryptography (Van Tassel,

[1] Oxford online dictionary: `http://www.oxforddictionaries.com`.

1969; Gardner, 1984; Bennett, 2004; Vatsa, Mohan, & Vatsa, 2012).

Table 2.1: Categorization of gibberish operators

Text	Sentence Level	Chomsky sentence "Colorless green ideas sleep furiously"
	Word Level	Jabber Wocky sentence, pseudo words – nonwords - logatomes, cryptography, scat singing
Speech	Cue Masking	Low-pass filtering, randomized content splicing, backward speech, sine-wave synthesis, spearcon synthesis
	Cue Manipulating	Utilization of speech synthesis technologies

The approaches that work directly on the speech signal can be generally categorized under two main strategies: *cue/content masking* and *cue/content manipulation via synthesis* (Banse & Scherer, 1996; K. Scherer, 2003). The core procedure in both of these approaches is to systematically manipulate or vary the acoustics cues so that the effect of paralinguistic aspects, such as vocal parameters, on emotions or on speaker attitudes can be studied (K. Scherer, 2003).

Cue/content masking approaches are usually applied on the speech signal and the verbal cues are masked, distorted or removed from these vocalizations. Some examples of this approach include: low-pass filtering (Friend, 2000; Knoll, Uther, & Costall, 2009; Snel & Cullen, 2013), randomized content-splicing (K. Scherer, 1971; Teshigawara, Amir, Amir, Wlosko, & Avivi, 2007), reiterant speech (Friend, 2000), backward speech (Johnson, Emde, Scherer, & Klinnert, 1986; K. Scherer & Ekman, 1982), sinewave synthesis (Remez et al., 1981; Remez & Rubin, 1993; Barker & Cooke, 1999) and more recently spearcon synthesis (Walker, Nance, & Lindsay, 2006; Palladino & Walker, 2007). In cue/content masking approaches each technique removes or distorts different types of acoustic cues from the speech signal while leaving others unchanged. This way, it is possible to study certain affective information carried by different vocal cues. However, in some of these masking techniques it is still possible to recognize the lexical content to a degree (Remez et al., 1981; K. Scherer, 1971; K. Scherer et al., 1972), such as in randomized splicing.

Cue manipulation via synthesis approaches basically utilize speech synthesis technologies to parameterize speech that then allows for systematic manipulation of the vocal parameters. As a result, the relative effect of these manipulation on people's affect interpretation can be studied (K. Scherer, 2003). An early example

of this approach is from (K. Scherer & Oshinsky, 1977), who used a MOOG synthesizer to create concatenated tones of sounds that were designed to resemble both sentence-like utterances as well as musical melodies, by specifically manipulating the pitch, rhythm, amplitude contour, timbre and tempo of tones. More recently, the use of speech synthesisers has become popular as reflected by the large number of publications on the subject (e.g., Cahn, 1990; Murray & Arnott, 1995, 1996; Burkhardt & Sendlmeier, 2000; Laukka, 2005; Schröder, 2001, 2003b; Schröder, Burkhardt, & Krstulovic, 2010). Although these techniques are not directly used to eliminate semantic information, semantically-anomalous pseudo-phrases such as Chomsky (1956) sentences, semantically neutral sentences (Burkhardt & Sendlmeier, 2000) or gibberish sentences (Oudeyer, 2003; Yilmazyildiz, 2006; Yilmazyildiz, Latacz, Mattheyses, & Verhelst, 2010) are synthesized in some studies to study emotion in speech without being affected by the linguistic content.

All these techniques had their places since very early years in the literature. Together with the advancements on robotic industry, attempts to utilize them in HRI studies have only been made in the last decade.

One of the earliest examples of gibberish-like speech developed for HRI purposes was from Breazeal (2000, 2002) in the form of babbling vocalizations for the robot Kismet. She created utterances as strings of various lengths by randomly combining predefined syllable structures, each containing random vowels and consonant phonemes. These strings were then synthesized with a commercial formant speech synthesizer (DECtalk). The parameters of the synthesizer (only the ones that have a global influence on the speech signal: voice quality, speech rate, pitch range, average pitch, intensity and the global pitch contour) were altered to convey affect, depending on Cahn's (Cahn, 1990) vocal affect parameter mapping. One single fixed mapping was used per each emotional quality of anger, disgust, fear, happiness, sadness, surprise and neutral. The system was then evaluated both by analyzing the acoustic features of the synthesized speech with respect to the acoustic correlates of emotion and by a perception experiment with human listeners. The acoustic analysis revealed a distinct profile for each emotion. Human listeners achieved about 60% overall recognition accuracy, with fear having the lowest recognition rate of 25%. The confusion occurred mostly between the emotion labels having similar characteristics such as fear and surprise (both share high arousal) and disgust, anger and sadness (sharing negative valence). These results revealed that confusions commonly seen with natural language also occurred similarly with gibberish speech.

Later, Oudeyer (2003) aimed at generating computationally inexpensive, exaggerated cartoon-like emotional speech by using simple algorithms and by controlling

as few parameters as possible. Oudeyer generated utterance strings that were composed of words with various lengths, but each containing a combination of open syllables of either CV or CCV type (C = consonant, V = vowel). As opposed to Breazeal, he synthesized the strings with the MBROLA speech synthesizer (Dutoit & Leich, 1993) based on a concatenative synthesis method. MBROLA received the following information: the strings to be synthesized, the duration and mean pitch of each phoneme and the stressed syllables. By altering mean & max pitch, pitch variation, mean duration, duration variation, accent probability, last word accent and contour values, this system was able to convey five emotional states: anger, calmness, comfort, happiness and sadness. The system was evaluated with a set of two listening experiments with human subjects: unsupervised and supervised tests (in the supervised test an example of each emotion with their labels was provided to the listeners prior to the test samples). The total recognition accuracy was about 57% for the unsupervised test and about 77% for the supervised test. Much better accuracies were achieved once the calmness/neutral affect category was removed in another experiment set (75% and 89% for unsupervised and supervised, respectively). In the unsupervised set, confusion occurred about valence for aroused emotion labels (between anger and happiness). It can be seen that when the users are supervised, utilizing examples of each affect category, eye catching improvements on decoding can be achieved.

Yet again, based on concatenative text-to-speech (TTS) solutions, Németh, Olaszy, and Csapó (2011) developed an auditory emotional and intentional state representation scheme. In this study they introduced the notion of "spemoticons (speech based emotions)" as an alternative to earcons (Blattner, Sumikawa, & Greenberg, 1989) and auditory icons (Gaver, 1986) in human-computer user interfaces. *Spemoticons* are obtained by modifying the intensity, pitch, and time structure of the Profivox TTS synthesizer (Olaszy et al., 2000) that uses diphones (CV, VC, VV, CC) and triphones (CVC) for speech generation. First, basic sound from a text-like character sequence is synthesized and then the prosodic modification is applied to have the final character of the sound, by the interactive TTS modification tool. The generated sound samples were evaluated with a perceptual test and the participants were asked to choose one of the possible 7 emotional categories for each sound sample: positive emotion, negative action, conflict, bad mood and its consequence, warning/anxiety, positive evaluation/commendation. After eliminating the most confused sounds, 9 samples were shortlisted as representatives of negative and positive categories. In this study no validation tests were performed that evaluate the perception of the spemoticons by human subjects.

Differently in (Yilmazyildiz, Mattheyses, Patsis, & Verhelst, 2006; Yilmazyildiz, 2006), a concatenative system based on a prosody transplantation technique was

developed to convey affect, in which the speech waveform was directly modified to to create gibberish speech. To convey the required affect, the system first selected a prosodic template from the database of the corresponding emotion. Then a carrier utterance having the same syllabic structure as the template was created from a neutral speech database by concatenating speech segments of various lengths. The pitch and timing structure of the expressive prosodic template was then copied on the neutral gibberish carrier utterance to produce the required affect. The listening test with human subjects revealed about 45% recognition accuracy for anger, fear, happiness and sadness. The biggest confusion occurred between anger and neutral, and also between fear and neutral. The lack of voice quality transplantation, together with the relatively poor quality of the database from which the prosodic templates were selected had a negative impact on the synthesized affect samples. When the listener was made familiar with the synthesized GS utterances (by presenting neutral GS samples), the synthesised emotions were recognised with more accuracy. This gives indications on a potential improvement in decoding an agents' emotion gradually with time.

All these studies have implemented and evaluated different methods and techniques of synthesizing affective GS, without evaluating how natural the generated synthetic speech was. Considering the importance of the naturalness of the generated speech, especially in social HRI, this was a gap. Addressing this, in Chapter 3 of this dissertation, the effectiveness of gibberish speech compared to semantically neutral or multiple levels of semantically charged speech in evoking the intended emotion was investigated, while also researching how natural the synthetic speech generated out of gibberish text strings would sound. Gibberish text was created by removing the semantics of an existing intelligible text in a certain language, instead of randomly creating it from scratch as did Breazeal and Oudeyer (Breazeal, 2000, 2002; Oudeyer, 2003). The semantics of an existing text in a language was destroyed by replacing the vowel nuclei of the text using a weighted swapping mechanism in accordance with their natural distribution in the same language. The generated gibberish text was then synthesized into speech by using TTS engines, which resulted in gibberish speech that resembles an unrecognized natural language, as intended. This result not only highlights the naturalness of the new gibberish language but also shows the potential of creating language specific gibberish. In the next step, the recognition of emotions from a gibberish language was experimented by synthesizing gibberish text with an open source emotional TTS system, "EmoSpeak", of Mary TTS synthesizer (Schröder, 2003b, 2003a) in comparison with supporting, confusing and neutral semantic samples. This was tested in a listening test with human subjects by using two emotion categories: happiness and sadness. It was found that with gibberish speech it is still possible to express emotions as effectively as with semantically neutral speech.

During these studies, it was highlighted that, once the gibberish text was fed into a TTS engine, the resulting affect strongly depended on the TTS engine quality. Also the voice quality of the emotions was not fully transmitted to the synthesized speech. This motivated the study in Chapter 4 of this thesis, to go for a data-driven method that starts with a high-quality emotion database. For this, various gibberish texts were created by replacing the vowel nuclei and consonant clusters of dialogue text scripts in accordance with their natural probability distribution in English and Dutch, respectively, and a voice actor portraying the basic emotions (anger, disgust, fear, happiness, sadness, surprise) and neutral speech, while vocalizing these gibberish text scripts, was recorded. Listening tests revealed about 81% recognition accuracy. This clearly high decoding accuracy, in comparison to earlier results in the field, strengthens the argument on the importance of voice quality and high-quality databases in affect expression. For generating more variations of the recorded gibberish utterances, a concatenative technique for swapping the units of an utterance with other units from the database of the related emotion was developedwhich will be described in Chapter 5 of this dissertation.

As with human-human interaction, the emotional information exchange with robots takes place on different layers of multi-modal interaction, in contrast to uni-modal interaction. Thus, the quality of the emotion decoding on the user side can be seen as a combined factor of the success in auditory, facial and gestural affect expression allowed by the affordances of the robots.

In respect to this, a multi-modal evaluation study that will be elaborated in Section 6.3 was performed. Specifically, it was investigated whether speech without semantic meaning can contribute positively or negatively to the emotion expression of robotic agents. Three layers of modalities are used in this study: auditory, facial and audio-visual. GS samples (conveying anger, disgust, fear, happiness, sadness and surprise) were thus either played alone (audio condition), or combined with facial expressions (audio-visual condition) in the robot Probo (Saldien, Goris, Yilmazyildiz, Verhelst, & Lefeber, 2008) to children subjects. It was found that children decoded the emotions of Probo from facial cues only with an accuracy of 42% and the usage of GS in combination with the facial expressions significantly increased the accuracy to 71%. This shows that GS helped resolving the ambiguities and confusions in facial expressions significantly. Later, in section 5.3 the relation between the voice characteristics and the robot morphology is also explored. Starting from the base utterances containing neutral, happiness and sadness gibberish speech samples, low and high-pitched samples were designed by time-scaling and resampling techniques that provided global spectral shifts. Subjects then evaluated the appropriateness of these utterances across two different robot platforms (Probo

- high volume animal-like green robot with a fur and Nao - low volume human-like gray robot with a plastic cover). The results show a clear relation between the robot morphology and the appropriate voice pitch. The high volume robot Probo was related more with the lower pitched voices while for the low volume Nao it were the higher pitched voices.

GS in multi-modal settings provides an important contribution to social HRI, as a step towards bringing the emotional expression much closer to the way it is seen in real-life. An additional step along the same line is employing GS in more real-life scenarios, such as in contextual settings, and evaluating whether there would be perceptual changes in interpreting GS.

In this respect, Chao and Thomaz (2013) utilized GS to evaluate another social aspect in HRI: turn taking during multi-modal interaction. They generated gibberish phrases by sampling random strings of phonemes each lasting 1–5 seconds in duration. The phrases were grouped by the prosodic endings expressing ellipsis, exclamation, interrogation, and statement. Their evaluation required subjects to interact with their robot Simon in a natural manner, and so they told the subjects that their task was to teach the robot about a variety of different objects. In reality, the robot did not do any learning, however; this request was used to evoke natural social behaviour from the subjects. In turn, the robot exhibited similar natural social cues, making both visual gestures and audible vocalisations whose timing and orchestration was influenced by the subject's behaviour. In order to avoid having to implement an NLP system (which, if it failed, could have had adverse consequences on interactions), they implemented a GS system in the robot in the same way as Breazeal (2002).

In Section 6.5 of this dissertation, another example of a real-life scenario design in a child-robot interaction study will be presented. Each child watched selected emotionally rich short animation movie clips together with the Nao robot, sitting next to each other. At the end of the clips, the robot communicated its emotion to the child in an affective gibberish speech and the child had to rate the recognized emotion on valence and arousal dimensions using the Self-Assessment Manikin (SAM) measurement tool (Bradley & Lang, 1994). There were two different robot profiles: having a congruent or a contradictory emotion with the dominant emotion in the movie clips and the children could mostly perceive the intended affective information from the gibberish speech generated by the robot.

2.3.2 Musical utterances

Together with the advancements in sound-producing capability of computing systems, the field of computer music has evolved significantly. The expressive nature of music has thus started being used as a more structural way of communication. Music has both simple and complex structures that can be utilized according to the needs. In the domain of human-computer interaction, there have been studies using music to debug software programs (Vickers & Alty, 2002) or to communicate graphical information for blind people (Alty, Rigas, & Vickers, 2005). Expressing affect is already in the nature of music. Music is basically coding the emotions, feelings and sensations of the composer by means of a musical score to be decoded by the listeners. Thus by modifying parameters such as pitch, rhythm, dynamics and timbre, affect can be encoded in the musical piece. This has similarities with expressing emotions with speech as discussed in (Juslin & Laukka, 2003) and it has even been hypothesized that music and speech has the same psychological evolutionary root (K. Scherer, 1995; Brown, 2000).

In a similar manner that Text-To-Speech technology has been utilised to create GS, some authors have drawn upon insights from musicology and musical theory to create utterance-like non-speech. Motivations for this come from the extensive body of research exploring how affect can be conveyed through music in general, as well as from the considerable overlap that exists between music and affective displays through the singing voice (see Juslin and Laukka (2003) for an extensive review of this).

Johannsen (2001, 2002, 2004) provides what is perhaps the earliest example of the use of MUs in a service robot to communicate intended directional motion trajectories (e.g. left, right, forward, backward), functional states (e.g. carrying a heavy load, waiting, low battery and near an obstacle) as well as the degree of urgency. The application scenario in this case was the use of a robot moving within a supermarket setting. Musical theory and notation was used as the basis upon which the utterances were designed. Subjects were tested for their comprehension and recall of learned sound/meaning associations. Following this, subjects were asked to draw out the trajectory of a simulated robot based upon the sounds that it made. In this example, the utility of the utterances has very much been focused on the communication of simple and very functional information about the robot (i.e. the robot is carrying something heavy, or the robot is about to turn left).

Besides using musical theory as a foundation for creating utterances, technologies developed for the music industry are also commonly used. For example, Jee et al. (E.-S. Jee, Kim, Park, & Lee, 2007; E.-S. Jee, Park, Kim, & Kobayashi, 2009) have

used musical notation, theory and computer-synthesisers to hand create a small collection of utterances that were designed to have a particular affective charge (happy, sad, fear and dislike) through the variation of acoustic features of tempo, key, pitch, melody, harmony, rhythm and volume. Subjects were then asked to perform three tasks to investigate the influence of the emotion stimulated by composed music and compared with the influence by robot's facial expression. Firstly, listen to each of the utterances and rate their emotion stimulated by composed music on a five point scale (very strong, strong, moderate, weak, never). Secondly, they were shown a cartoon face of a robot with an expression matching each of the labels, and again were asked to rate their stimulated emotion on a five point scale. Finally, the subjects were presented with both the face and utterance for a given label and asked to rate their feelings on a five point scale. Their results showed that both the utterances and facial expressions alone 40% - 70% of the subjects reported strong happy, sad, fear or dislike, while when combined together, the report of the strong feeling increased to 80% -90%. This shows that by combining the two modalities, the subjects feel a more intense emotion than when the face and utterances were presented individually, a finding that is in agreement with those of the multi-modal study which will be discussed in Section 6.3.

E. Jee, Jeong, Kim, and Kobayashi (2010) furthered this work by creating five sounds that were designed to convey particular intentions (affirmation, denial, encouragement, introduction, question), and three emotions (happy, sad, shy), again using musical theory and computer synthesisers to change the intonation, pitch range and timbre of the utterances. Subjects were then presented with each of the utterances and again asked to rate their feeling on a five point scale (very strong, strong, moderate, weak, never). Their results showed that more than 55% of the subjects reported that the musical sounds developed were sufficient to express intended intentions and more than 8O% of the subjects thought that the developed sounds were sufficient to express intended emotions. That said, what these examples do demonstrate is that using this music inspired approach also has potential utility.

Ayesh (2006, 2009) took inspiration and insights from the world of musicology but also from natural language. He developed an algorithm that allows a robot to create, on the fly, an artificial "language" for emotional interaction. He created a syntactical definition of a musical language (Musical Language for Emotional Interaction between Robots - MLEIR) that is capable of communicating emotions for animal-like robots. This was implemented on a Lego robot to express urgency, stress, excitement, calm and fear based on a pet dog's observed behaviours. He argued that the mixture of these feelings leads to portrayals of different emotional states (for example, an excited and stressed robot demonstrates the feeling of anger, or the combination of stress and urgency crystallize into fear). As no human-based

evaluations were carried out, the system's ability to accurately convey a desired affective colouring is unclear.

Similarly, Esnaola and Smithers (2005) also developed a "language" based upon musical theory, which can be whistled or played on simple musical instruments. They chose to use frequencies out of the human speech and singing ranges (between 1046.52 Hz (note C) and 2637 Hz (note E)) to avoid confusion when people are talking and communicating with the robot. They created an alphabet out of 10 musical notes from which then they created words (noun, verb, adverb and interjections) with pre-defined meanings that then constituted phrases. The musical language requires to learn a melody or to use an interface (PDA, mobile phone) to generate it, which can be considered as cumbersome. However, they argue that the learning of a melody can be performed naturally by humans. Furthermore, they state that with only a little learning, recognition and thus communication is much more robust than the use of spoken natural languages in noisy environments.

2.3.3 Non-linguistic utterances

Non-speech like sounds have been used as a means of feedback during human-machine interaction for quite some time now. For example, efforts in conveying information through sound dates back to the 19th century with the inventions of telephone in 1876, phonograph in 1877 and radiotelegraphy in 1895. With these inventions, transformation of sound waves into electric signals, and vice versa, came to life (Dombois & Eckel, 2011) which has then driven the techniques and technologies to register and display sound. Years later the terms that are used today to describe the information representation started to be introduced one after another: sonification, auditory icons, earcons, and now, Non-Linguistic Utterances.

Sonification is the technique of rendering sound in response to data and interactions (Hermann, Hunt, & Neuhoff, 2011). This technique makes use of non-speech sounds/audio to convey information from non-audio mediums into the audio medium. It is primarily used as a means to perceptualise an input data stream. A prominent example of this is a Geiger counter, which conveys the level of radio activity present in an environment through the frequency of clicking.

Auditory Icons (Gaver, 1986) are sounds that are designed to resemble the sounds associated with real-world objects and actions. Thus, they have a semantic connection to the physical events that they represent, and as such may be considered as auditory representations of "visual" icons. They are created by mapping events that occur in a computer-based world into events occurring in the real world. Gaver (1986) proposed three categories of auditory icon based on their degree

of arbitrariness: *symbolic, metaphorical* and *nomic*. *Symbolic icons* draw their meaning from social conventions (e.g. the siren of an ambulance) and thus can be difficult to learn if one is not aware of the social conventions. *Metaphoric icons* hold strong similarities with the action of events they represent (e.g. a falling pitch is associated with a reduction in the size or altitude of an object). Finally, *nomic icons* are based upon sounds that are created due to physical causation (e.g. the sound of a paper being thrown into a rubbish bin when deleting files on a computer system).

Earcons (Blattner et al., 1989) are brief, structured and abstract synthetic sound patterns that are used in many modern technologies such as computers, games consoles and smart-phones to represent a specific item, event, meaning, state or label. For example, the sounds associated with a computer starting up or shutting down, or when an error occurs. Due to the abstract nature of these sounds, their relation to their meaning is something that must be learnt by users, which is one of the drawbacks in comparison to Auditory Icons (Walker et al., 2006; Dingler, Lindsay, & Walker, 2008).

While these three strategies for acoustic communication are now commonplace in modern day Human-Computer Interaction, we draw an important distinction between them and NLUs, with this being their use as *social cues*.

Similarly to GS, the roots of NLUs are rather old and are certainly not in the world of robotics. NLUs, like GS, have a small history in the field of psychology, where they have also been used as a means to explore how affect may be communicated through acoustic signals. For example, K. Scherer et al. (1972) used tones to produce acoustic utterances to systematically explore what minimal cues were needed within an acoustic signal to convey affect. However, as technological developments in speech synthesis technologies improved, the use of such NLUs in this respect has ultimately been limited. On the other hand, the emergence of the social robot has lead to the desire to display softer elements of social interactions, such as positive or negative attitudes and affect.

For example, Komatsu (2005) explored how utterances can be designed in order to convey either positive or negative attitudes, and agreement or disagreement, on the robot's part. It was found that utterances with rising frequency modulations were commonly rated as positive or expressing agreement, while utterances with a falling frequency modulation conveyed a negative attitude. These can be considered as very *iconic* sounds (e.g. earcons) as similar types of sound are commonly used in everyday technologies such as mobile phones, computer programs and even computer games, as a means to provide feedback on whether something positive or negative has happened.

Read and Belpaeme (2012) investigated how young children interpret NLUs, which had no musical influence or inspiration. In their experiment, children listened to a broad variety of NLUs and affectively rated them using a facial gesture tool, the AffectButton (Broekens & Brinkman, 2013). Utterance synthesis was achieved using the free, open-source real-time computer music synthesizer, SuperCollider (McCartney, 2002). Similarly to K. Scherer et al. (1972), the goal was to explore if particular prosodic arrangements were more robust at conveying affect than others. Their results revealed interesting insights, namely that children readily see affect in NLUs made by robots, and that their interpretations are not subtle. In their own words, the robot is either *happy, sad, scared* or *angry*. Moreover, they found that children were not coherent in their interpretations: children provided different interpretations when rating the same utterances.

Following on from these findings, Read and Belpaeme (2013) have found that adults' affective interpretations of NLUs are subject to Categorical Perception. The results of their experiment conducted with adults revealed that indeed affective interpretations of NLUs were drawn to basic affect prototypes, showing that subtle changes in acoustic properties of NLU did not result in subtle changes in affective interpretation. While other studies have been focused upon which different affective states SFUs can be portrayed to people, this study was focused upon understanding what the landscape of the transition between different affective interpretations looks like.

Schwent and Arras (2014) have taken this further with their robot Daryl. They developed an architecture for sonic human-robot interaction that allows robotic NLUs to be generated and synthesized in real time as the robot interacts with the environment. Also using SuperCollider they create complex utterances inspired by natural language comprised of simple sounds that are concatenated in a hierarchical manner that is analogous to the structure of natural language. With this architecture they explored how the physical social cues such as head, body and ear pose may be aligned with the parameters of their utterances in order to convey a particularly affective state convincingly. Furthermore, they show that the modulation of utterances can be controlled through the robot's perceptual inputs. Specifically they have used information regarding inter-personal distance between a human and Daryl to produce real-time, reactive sonic feedback on peoples' proximity to the robot.

Recognising that the acoustic properties of utterances is not the only thing that can impact how people perceive and interpret robotic utterances, there have also been efforts directed at exploring the tri-directional relationship between

utterances made by a robot, the type of robot embodiment, and peoples' interpretation of these utterances. As robots come in many shapes and sizes it has been deemed important to understand that there is an important relationship between the physical appearance of a robot, and the auditory behaviour and characteristics that it exhibits.

Komatsu and Yamada (2007, 2008, 2011) have investigated how different agent embodiments would impact how the same utterances were interpreted. Utterances consisting of simple tones with either increasing or decreasing pitch were embodied in a PC, an Aibo robot, and a mobile robot made of Lego. In their experiment, subjects were presented with each of the three embodiments, and asked to rate how positive or negative they thought the utterances were. Their results showed that when the utterances were made by the PC, people showed a high degree of accuracy in interpreting the utterances, while this was not the case when the utterances were made by the Aibo and Lego robots. More specifically, they found that subjects struggled to correctly identify the positive utterances as positive, while their identification of the negative utterances remained high.

Read and Belpaeme (2010) investigated whether the morphology of a robot impacts how appropriate the utterances it makes are. They conducted an experiment where both a humanoid robot and an animal like robot made different types of utterances (human-like, animal-like and NLUs). Subjects were asked to indicate whether they thought that the different robot-utterance pairs were appropriate. They found that people preferred it when a humanoid robot makes human-like utterances, over animal-like utterances, and visa-versa for the animal-like robot. Furthermore, they found that NLUs were deemed as an acceptable utterance for the humanoid robot to make.

Different embodiments afford different forms of visual affective displays. For example, humanoid robots are able to make gestural displays with their limbs (e.g. Beck et al., 2013), robots with actuated (Breazeal, 2002; Ribeiro & Paiva, 2012; Trovato et al., 2013) or projected (e.g. Delaunay, de Greeff, & Belpaeme, 2010) faces are able to produce facial gestures with varying degrees of complexity and even simple wheeled robots can convey affect through attached limbs such as tails (Singh & Young, 2012) and through locomotion alone (Saerbeck & Bartneck, 2010). Such robots afford multi-modal interaction, and as such, this can introduce the complexity of behaviour synchronization: when to make affective displays, and how to synchronize these across different modalities.

Bramas, Kim, and Kwon (2008) sought to investigate this and developed a sound system to be used in synchrony with the main behaviour generation in order

to increase the impact of gestures made by a robot. Their system modified acoustic properties such as the volume, and prosody of utterances as well as applying effects such as echo and flangers. Rather than gearing their system toward conveying particular affective states, their system focuses upon the problem of when a robot should make utterances, and how this should be synchronized with other modalities such as gestures to emphasize emotional and gestural impact.

As with human-human Interaction, social cues used by robots should not be studied and explored as a single modality or in isolation. There are many factors that impact how people interpret social cues, such as the context and situation in which they are used, which in turn can also impact how people behave as a result. This has also been an emerging aspect of HRI in general that has begun to receive attention.

Komatsu et al. (Komatsu, Yamada, Kobayashi, Funakoshi, & Nakano, 2010) explored how NLUs used by a robot can bias how a person performed a task. More specifically, the setup involved having a subject play a treasure hunting game on a computer. The game showed a strait road, with hills appearing along the way. Under one of the hills a golden coin was hidden, and subjects had to guess under which one. Sitting next to the subjects was a Lego robot that told subjects which hill the coin was under, and then made an utterance with either a flat pitch contour or a falling pitch contour as a means of indicating how confident the robot was in its predication. Their results showed that when the robot's predication was accompanied by an utterance with a falling pitch contour, they rejected the predication significantly more than when an utterance with a flat pitch contour was used. In essence, the pitch modulation had a direct impact over the perception of how confident the robot was about the information that it gave.

Extending this work into communicating the level of confidence that an agent has about information that it presents to people, Komatsu and Kobayashi (2012) conducted a further experiment to see whether the use of NLUs can mitigate the potential adverse effects that the presentation of incorrect information may have. Comparing NLUs and natural language, their results show that when the computer provided completely correct information, natural language was preferred over the use of NLUs. However, in situations where the agent's confidence in the information that it was providing was misjudged, and thus the agent was shown to be confident about information that was ultimately incorrect, NLUs were the preferred method of expression regarding the agents confidence. Their argument in this work is that currently computers and robots are not perfect - they make mistakes, and that when agents use natural language to communicate, this sets a high expectation level, and when this expectation is violated, this evokes an adverse reaction in people. This

is a tangible example of how SFUs may be used to manage the expectations that people have of robots.

With respect to exploring how situational context impacts the perception of NLUs, Read and Belpaeme (2014b) conducted a video based experiment where subjects were shown videos where a robot was subject to an action (slapped, flicked in the face, having its eyes covered, being stroked on the head, and being kissed on the head), and the robot either made no utterance, a positive NLU or a negative NLU. Subjects were asked to view each video and provide a rating as to how they thought the robot felt based upon what happened in the video. The results showed that subjects' interpretations were heavily biased by the context in which they were made. Put plainly, two different NLUs had the same interpretation when made within a given context, and the same NLU would yield different interpretations when presented in different contexts. Their results also showed a subtle effect whereby when the perceived valence of the action/context was aligned with the perceived valence of the NLU, subjects' interpretations were more intense and extreme than when the NLU and context were misaligned.

2.3.4 Paralinguistic utterances

The first use of "paralinguistics" came in the mid last century restricted in the human-human communication domain. Since then, the definition and the subject area varied among the researchers. In very broad definition, "paralinguistic/paralanguage" corresponds to the study of vocal (beyond verbal message or speech) and non-vocal signals (gestures, postures, etc.) and in a narrow definition the non-vocal signals are excluded (Schuller & Batliner, 2014). The examples of broad definition include body language, gestures, facial expressions and the vocal factors of speech while the examples of the narrow definition include only the vocal factors of speech. The paralinguistic vocal factors can come as *modulated/embedded onto the linguistic chains* (such as pitch and voice quality) or as *stand-alone vocal events* (such as filled pauses and vocal outbursts).

The *modulated/embedded* form of paralinguistic is already at the core of especially Gibberish Speech as the major part of affect expression is realized with such aspects of paralinguistic in GS. This also gave inspirations to NLUs and MUs: as such they all integrate some form of paralinguistic aspects (such as pitch) in their nature. Thus, modulated/embedded form of paralinguistic is not the subject of this section as they are already dealt with implicitly (and sometimes explicitly) in earlier SFU types.

Stand-alone forms on the other hand, have not been covered yet and will be

the focus of this section as another SFU category in HRI. In essence, Paralinguistic Utterances (PUs) in this section stands for the stand-alone paralinguistic forms.

In human-robot/agent interaction, the amount of studies regarding PUs in this respect is very limited. The specific types of PUs in these studies involve back-channel signals (Ward, 1996), pause fillers and affect bursts (K. R. Scherer, 1994). Schröder (2003a) provided a detailed experimental study of affect bursts and showed that affect bursts are highly recognizable (81% when presented as isolated audio) and can convey a number of different emotion categories reliably. He envisioned that in the medium term, the use of PUs in this respect might become interesting for application in emotion expression in the field of conversational agents and indeed it has been getting more attention recently (Schuller & Batliner, 2014).

For example, Prendinger (Prendinger, Becker, & Ishizuka, 2006) investigated the impact of a virtual agent's affective and empathetic behaviour on the user-side in a game setup. PUs in the form of grunts and moans were used as the vocal modality supporting the emotions in the facial expressions. Their results suggested more general findings about affective interaction, as they found that the absence of the agent's display of negative emotions is conceived as arousing or stress-inducing.

Becker-Asano et al. (Becker-Asano & Ishiguro, 2009; Becker-Asano, Kanda, Ishi, & Ishiguro, 2011) studied one certain type of PUs which is "laughter". They explored the naturalness of various laughter for two humanoid robots with and without situational context. The laughter which seemed to fit both robots in a context (in response to a joke), was not found to be fitting any of the robots without a context. This is yet another example on the importance of context in HRI and in SFUs. Additionally, the morphology of the robots on the perceived naturalness of the samples didn't show any major effect. They speculate that any real differences for morphologies could probably be dominated by the major differences in the laughter samples themselves. Of course laughter is a very specific and distinct type of PUs and as Trouvain and Schröder (2004) suggest it requires a careful design/modelling, especially on the intensity, for an intended level of social bounding effect.

More recently, Schröder et al. (2012) utilized PUs in the form of backchannel signals on the sensitive artificial listener (SAL) system. With SAL the aim was to build an autonomous social conversational agent that focuses on emotional and non-verbal behaviour (both gestural and vocal) and thus reduces the need for spoken language understanding, task modelling, etc. The SAL system analyses the non-verbal cues of the interaction partner and replies back using back-channel signals such as "huh, wow", smiles and head nods along with some predefined phrase scripts. The results revealed that SAL with expressive backchannel cues caused the

interaction partner to show more behavioural engagement. Similarly, Kobayashi and Fujie (2013) also implemented a conversation protocol utilizing back-channel paralinguistic information for human-robot interaction.

It should be noted that studies regarding PUs are in the stage of exploring their potential and are not (yet) intended to be utilized in vocal/sonic synthesizers (yet there exists some exceptions such as (Niewiadomski et al., 2013)). Moreover, their usage with robotic agents is very limited. For these reasons, PU studies will not be included in the evaluations and analysis in the rest of this chapter.

2.4 Discussion

2.4.1 Summary, evaluations and discussion

Given the above review, it is possible to draw out a number of different observations and insights regarding the current state of SFU research. This section seeks to provide a comprehensive summary of the review. To aid in this process, Table 2.4.1 lists all the studies that have been reviewed, outlining a number of different aspects of the studies that are of interest and hold importance. The studies have been reviewed in detail and outlined in terms of authors, publication year, SFU method utilized (GS/MU/NLU), parameters modified, emotions portrayed, and the evaluation related aspects such as affective metric that measured the affect recognition, sample size, participant profile (number of subjects, age range, culture), display medium (isolated audio/robot pictures-videos/embodied robot) and recognition accuracy where applicable.

Display Medium and Embodiment:

While all the studies concern the application domain of social robots, many of the works reviewed have not employed an actual robotic agent, physical or virtual, as the display medium during the evaluation process. The issue of whether utterances are embodied in some agent has important implications on the generality of studies. Komatsu and Yamada (2007), Read and Belpaeme (2010) as well as the results in Section 5.3.2 of this thesis, have found that the morphology of a robot's embodiment has a strong influence over peoples' perception of SFUs both with respect to the inferred affective content as well as whether utterances are deemed "appropriate" for a given morphology. Also, scenarios with embodied robots showed more interaction (Fridin & Belokopytov, 2014), increased empathy (Seo, Geiskkovitch, Nakane, King, & Young, 2015) and enjoyment (Leite, Pereira, Martinho, & Paiva, 2008) from the users. Furthermore, there are important proxemics aspects that need to be considered also as users are sensitive to these (c.f. Mumm & Mutlu, 2011; Rae,

Takayama, & Mutlu, 2013; Takayama & Pantofaru, 2009).

The issue of embodiment is essentially a double-edged sword however. While studies that do not use a real robot as the display medium are able to study peoples' perception of expressive displays such as SFUs without the bias of embodiment, in the long run, it is unclear whether their findings still hold true when utterances do become embodied. There is a difficulty here however, as robots come in all shapes and sizes, it is unclear how robots should be characterized and thus the different embodiments, and the relation to acoustic behaviour be accounted for and represented quantitatively.

Affective Portrayals and Parameters Modified:

Concerning the affective portrayals, there is no standardization and most of the time the affects studied are driven by the intended applications of the robots in question. The majority of the studies worked on *categorical portrayals* (e.g. happiness, anger, sadness, etc.), while a few of them focused on broader terms such as *positive/negative* emotions or used concepts of *affective dimensions* (Cowie & Cornelius, 2003). The portrayals are achieved through modifying various sound parameters, which ranges from altering a single parameter to utilizing natural alterations in human speech production. Among the palette of parameters modified, pitch and duration were the two most common parameters that were altered almost in every study to affectively charge SFUs. However, when this occurs, an important assumption is being made: that the utterance is indeed charged in such a manner that it induces the desired interpretation in the listener.

However, the validity of the method of designing SFUs for particular affective portrayals can be questioned. In some circumstances this is not a valid assumption to make, as highlighted by Read and Belpaeme (2010). Crucially this requires prior knowledge and valid information regarding the mapping between the acoustic features of an utterance and the affective interpretation. Where this information exists, it becomes possible to create an utterance with confidence that it will induce a particular interpretation. For example, where already tried and tested technologies such as TTS are used to create affect-laden GS.

In studies in which new synthesizers are evaluated (e.g. Schwent & Arras, 2014) it is not possible to design utterances with a particular affective meaning in mind as there is no validated affective mapping between utterance features and interpretation. In these cases it is necessary to explore the parameter space of the synthesizer with the aim to uncover this mapping (such as in Read and Belpaeme (2012)). However, the drawback is that this requires negotiating an enormously

large parameter space that is a challenging problem in itself.

Affective Metrics:

Among the metrics used to evaluate the SFUs, generally either *forced choice* or a form of *continuous scale measurement* tools have been utilized. There are various advantages and disadvantages on the usage of both metrics.

In most cases when categories are used (in *forced choice measurements*), subjects are asked to select one of the emotion labels[2] that best matches their interpretation of the utterance. While this selection of labels is simple, intuitive and relatable to a broad range of subjects from different age groups and cultures, it does not provide a high resolution with regard to understanding subtle acoustic differences between utterances such as how intense an emotion is.

Continuous scale assessments have been utilized in even more different forms. These differences sometimes occurred in the affective dimensions assessed (one or more affective dimensions of *valence, arousal, activation* used in different studies) or in the tools that have been employed (some used picture based assessment such as the Self-Assessment Manikin (SAM) or Affect Button while some others utilized numeric Likert Scale). The main advantage of using continuous metrics is the higher resolution that they afford – the data captured is essentially *richer* in comparison to nominal data captured from affective categories. Moreover, continuous scales also lend themselves to machine learning, which is particularly useful when attempting to learn a mapping between a desired affective interpretation and the parameters of an utterance. The main drawback however is that affective dimension and continuous scales are often difficult to explain to naïve subjects, due to the general abstract nature of the dimensions (Broekens & Brinkman, 2013) .

Apart from these two main evaluation metrics (forced choice and dimensional measurements), a small number of studies have employed *metrics based on observable behavioural change* in the subjects. Although this type of assessment is harder to implement and process (for example, the cumbersome process of video coding often required to process the data collected), it is encouraged in HRI (and is gaining more traction as an acceptable method) as it measures the subjects without letting them being aware of that they are assessed and allows interactions to unfold without the need to interject interactions in order to perform explicit measurements. This also provides an experimental setting that is more akin to real-world HRI and thus increases the ecological validity of the evaluation.

[2]Usually the "basic emotions" (Ekman & Friesen, 1971) but also varies depending on the application intentions of the robots in question.

Table 2.2: Summary of studies on SFU included in the review

Study	SFU Cat.	Parameters modified	Emotions Portrayed	Affective metric	Sample size	Nb. of Subjects	Age range	Cross/ within-cultural	Display medium	Recognition accuracy
Breazeal (2001)	GS	Pitch-related parameters: accent shape, average pitch, contour slope, final lowering, pitch range, pitch reference line. timing-related parameters: speech rate, stress frequency. voice quality parameters: loudness, brilliance, breathiness laryngealization, pitch discontinuity, pause discontinuity. articulation parameter: precision	Anger, disgust, fear, happiness, sadness, sorrow	Categorical	18	9	23-54	NR	Audio clips	60%
Chao and Thomaz (2013)	GS	Strings of phonemes with different prosodic endings	Exclamation, interrogation, statement, ellipsis, extraverted, intraverted	NA	NA	30	17-45	NR	Embodided robot	NA
Nemeth et al. (2011)	GS	Intensity, duration and pitch	Positive emotion, negative action, conflict, bad mood, anxiety/warning, commendation/positive evaluation	Categorical	44	54	NR	NR	Audio clips	NA
Oudeyer (2003)	GS	Pitch contour, duration, overall volume	Anger, calm, comfort, joy, sadness	Categorical	30	8	Adults	Multi-cultural	Audio clips	57 % (unspurervised), 75 % (supervised)
Wang et al. (2014)	GS	Acted emotions	Anger, disgust, fear, happiness, sadness, surprised	Dimensional (SAM tool)	12	10	7-9	NR	Embodied robot	NR

Table 2.2 (continued)

Study	SFU Cat.	Parameters modified	Emotions Portrayed	Affective metric	Sample size	Nb. of Subjects	Age range	Cross/ within-cultural	Display medium	Recognition accuracy
Yilmazyildiz (2006a, 2006b)	GS	Pitch, duration, volume/energy	Anger, fear, happiness, sadness	Categorical	16	7	Avg. 27	Multi-cultural	Audio clips	45%
Yilmazyildiz et al. (2010)	GS	Pitch, range, (pitch, range) dynamics, accent (prominence, shape, slope), boundary type, rate, nmb of pauses, (pause, vowel, nasal, plosive, fricative) duration, volume	Happiness, sadness	Categorical	10	9	26-37	Multi-cultural	Audio clips	61% (gibberish)
Yilmazyildiz et al. (2011)	GS	Acted emotions	Anger, disgust, fear, happiness, sadness, surprised, neutral	Categorical	112	10	27-32	Multi-cultural	Audio clips	81%
Yilmazyildiz et al. (2013a)	GS	Acted emotions	Anger, disgust, fear, happiness, sadness, surprised	Categorical	27	35	10-14	Within-cultural	Audio clips and robot video clips	42% (visual), 65% (audio), 71% (audiovisual)
Yilmazyildiz et al. (2013b)	GS	Acted emotions, global spectral and vocal tract modifications	Neutral, happiness, sadness	NA	18	8	28-33	Multi-cultural	Robot video clips	NA
Ayesh (2007)	MU	Pitch, duration, volume and sound count.	Happiness, sadness, fear, disgust and surprise	NA	NA	NA	NA	NA	NA	NA
Ayesh (2009)	MU	Pitch, duration, volume and sound count.	Urgency, stress, calm, excited, fear of unknown, fear from memory.	NA	NA	NA	NA	NA	NA	NA
Esnaola and Smithers (2005)	MU	Musical note	NA	NA	NA	NA	NA	NA	Embodied robot	NA
Jee et al (2007)	MU	Tempo, key, pitch, melody, harmony and rhythm	Happy, sad, fear, dislike	Categorical	4	20	Adults	Within-cultural	Audio clips and cartoon face images	53.7% (visual), 65% (audio), 85% (audiovisual)

Table 2.2 (continued)

Study	SFU Cat.	Parameters modified	Emotions Portrayed	Affective metric	Sample size	Nb. of Subjects	Age range	Cross/ within-cultural	Display medium	Recognition accuracy
Jee et al (2009)	MU	Tempo, key, pitch and volume	Expectation, dislike, pride, anger	NA	4	NA	NA	NA	NA	NA
Jee et al (2010)	MU	Intonation, pitch and timbre	Affirmation, denial, encouragement, introduction, question, happiness, sadness	Categorical	6	20	20-25	Within-cultural	Embodided robot	55% (intentional), 80% (emotional)
Johannsen (2001, 2002, 2004)	MU	Sound color(with different musical instrument), rhtym and melody (time and frequency editing)	Urgency, expressiveness, annoyance for directions (left, right, etc.), robot states (heavy load, waiting, near obstacle, low battery)	Likert scale on urgency, expressiveness and annoyance	32	8	NR	NR	Audio clips	NA
Bramas (2008)	NLU	Volume, pitch and tempo, sound effects (such as echo and position effects)	Exciting, sadness/sick, hurry, lazy, happy, very happy, neutral	NA	NA	NA	NA	NA	NA	NA
Komatsu (2005)	NLU	Pitch, duration	NA	Categorical	44	23	20-42	Within-cultural	Audio clips	NA
Komatsu (2010)	NLU	Pitch	High confidence, low confidence	Behavioural	2	19	22-25	Within-cultural	Embodied robot	NA
Komatsu (2011)	NLU	Pitch	High confidence, low confidence	Behavioural	3	20	21-23	Within-cultural	Embodied robot	NA
Komatsu (2012)	NLU	Pitch, duration, prosody, speech rate	High confidence, low confidence	Behavioural	6	20	21-28	Within-cultural	Embodied robot	NA
Komatsu and Yamada (2007)	NLU	Pitch	Positive, negative, undistinguishable	Categorical	8	9	21-24	Within-cultural	Embodied robot and PC	40% (Lego robot), 36% (Aibo robot), 80% (PC embodiment)

Table 2.2 (continued)

Study	SFU Cat.	Parameters modified	Emotions Portrayed	Affective metric	Sample size	Nb. of Subjects	Age range	Cross/ within-cultural	Display medium	Recognition accuracy
Komatsu and Yamada (2008)	NLU	Pitch, duration	Positive, negative, undistinguishable	Categorical	8	20	19-24	Within-cultural	Embodied robot and PC	68% (Lego robot), 65% (Aibo robot), 82% (PC embodiment)
Read (2010)	NLU	Pitch, pitch range, pitch contour, speech rate, pause-ratio, sound unit count, timbre	NA	Categorical	20	61	Adults	Within-cultural	Robot image with an audio clip	NA
Read (2012)	NLU	Pitch, pitch range, pitch contour, speech rate, pause-ratio, sound unit count, carrier wave	NA	Dimensional (Affect-Button)	90	42	6-8	Within-cultural	Embodied robot	NA
Read (2013)	NLU	Pitch, pitch range, pitch contour, speech rate, pause-ratio, sound unit count, carrier wave	NA	Dimensional (Affect-Button)	6	28	Adults	Within-cultural	Embodied robot	NA
Read (2014)	NLU	Pitch, pitch range, pitch contour, speech rate, pause-ratio, sound unit count, carrier wave	NA	Valence Likert Scale	25	301	Adults	Within-cultural	Robot video clips	NA
Schwent and Arras (2014)	NLU	Pitch, pitch range, prosody, carrier wave	NA	NA	NA	NA	NA	NA	Embodied robot	NA

Note. Types of SFU category includes Gibberish Speech (GS), Musical Utterances (MU) and Non-linguistic Utterances (NLU). Avg. = Average, NA = Not applicable, NR = Not reported (means that no information was provided).

Subjects & Stimuli:

The sample size and the number of subjects evaluating these samples also vary a lot across the studies. Generally, very few utterance samples/stimuli have been evaluated in subjective tests and sometimes even just one utterance per emotion has been tested. This partially depends on the number of emotions studied. When the number increases, the size of the test also increases, which makes the test harder to complete for the subjects as they have a limited attention span. Thus, the sample size usually is reduced as a solution. However, the major drawback of this is that it reduces the degree to which results can be generalized.

The subject sample size used within a study has also been quite varied, ranging from roughly 10 to over 200, with an average of 33. This too leads to issues surrounding the generalisation of results. Studies that have reported larger subject sample sizes were online studies that have utilised crowdsourcing methods for HRI studies (Breazeal, Depalma, Orkin, & Chernova, 2013), but these too come with their own set of drawbacks.

Recognition Accuracy:

The recognition accuracies, which were mainly derived from the forced choice metrics, varied between 45% and 81%. This wide range is mainly due to the techniques used and parameters modified to produce affect-laden utterances of course. However, the differences in utilization of the forced choice metrics in measuring affect also have an influence. For example, the inclusion of "I don't know", "none" or "neutral/calm" option in the choice alternatives has been shown to have a big impact on the recognition rates (Oudeyer, 2003). The advantage of having this option is that it reduces the noise in the data in case none of the other answers apply to the user as well as that it reduces the pressure to give substantive responses felt by respondents who have no true opinions. However, the pitfall is that no-opinion options may discourage some respondents from doing the cognitive work necessary to report the true opinions they do have and thus may prevent the measurement of some meaningful opinions (Krosnick et al., 2002).

Focus Group & Culture:

It is also noteworthy that the majority of studies have focused on adults for evaluation and only a select few concerned themselves with child-robot interaction. Employing children has notable advantages in HRI. Children do not see robots as mechanical machines, and they readily anthropomorphise robots and maintain the illusion that they have life-like characteristics (Belpaeme et al., 2013). This has the advantage that new robots (or "creatures") with new physical forms and novel

vocalizations are likely more acceptable in their imaginary world than with adults. Despite of these potential advantages, the reactions of the children to SFUs are not fully discovered yet.

Moreover, Child-Robot Interaction is an area that is receiving considerable attention from both academia and industry as it provides an easy entry point for real world applications for social robotics. As such, there is a very real opportunity to endow these robots with SFUs and see their use in a wide spectrum of applications. However, it is vital to understand how the different age groups (and genders) respond to SFUs in order for their potential to be fully harnessed.

It is also noted that the majority of the studies were performed within cultural settings. Although some of the evaluations were performed with participants coming from multi-cultural and multi-language backgrounds, no real cross-cultural analysis (such as in Abelin and Allwood (2000) and Tickle (2000)) have taken place. As stated before, a potential advantage of SFUs is the fact that they are not bound to a single language or culture. In natural language, the cultural dependencies might play an important role in decoding the intentions and emotions (Shochi, Aubergé, & Rilliard, 2006; Mac, Aubergé, Rilliard, & Castelli, 2010; Burkhardt et al., 2006). However, it is unknown whether the interpretation of SFUs is similar across different cultures. Yet Tickle (2000) have investigated common tendencies in the acoustical correlates of basic emotions across different cultures (American and Japanese) by using GS samples and found very small differences which stands as first indications towards interpretation that GS may indeed not be bounded to a particular culture in carrying the emotions. However, this is only between two cultures and needs to be extended. This cross-cultural aspect is an important issue to address for SFUs, as it will clarify to what degree the already existing research may be generalized to and utilised in different cultural settings.

2.4.2 Grand challenges, future directions and discussion

In this section, the current "grand challenges" for SFUs are highlighted and discussed, while some initial approaches are proposed that may be adopted to begin addressing these challenges. This field can be considered to be very much in the stages of infancy when compared with other areas of HRI. Thus, the motivations in this section are to outline important aspects that require attention in order for SFUs to become more integrated with more mainstream HRI research.

Until recently SFU research, recursively went through cycles of the continuous development of new methods and algorithms for creating and synthesizing utterances, and the need to validate these new methods with respect to their ability to

convey different affective states. The development of new techniques and methods is certainly beneficial for the field as it keeps in touch with the developments in the fields that feed SFU research (e.g. Speech Synthesis). Moreover, it also allows further exploration to the "sound scape" that SFUs can encompass, as the scope for different "voices" that SFUs can provide is enormous.

However, a constant cycle of development results in a lack of attention toward understanding other aspects of SFUs, which are very important for their use in broader HRI. As such, it can be considered essential that the efforts into exploring new methods of synthesizing SFUs go further than just utterance creation, but also explore their use in scenarios that are more representative of real world HRI. Essentially, the exploration of the large soundscape holds little value if these other important aspects of SFUs are not considered. Furthermore, because each synthesizer is different, it is difficult to assess whether the results of evaluations conducted with one synthesizer are general enough to be used to inform the design of utterances using a different one. Thus, comparison between synthesizers (somewhat similar to Blizzard Challenge[3] for speech synthesizers) is required, and as the number of these grows, this becomes an ever more challenging task.

An extension to the idea of studying SFUs in ways that are representative of real-world HRI is studying SFUs not as a single modality. Social interaction between social robots and humans takes place through multi-modal interaction, not uni-modal interaction. As such, studying SFUs in a uni-modal manner is likely to produce results and insights that may not hold true when SFUs are used during multi-modal interaction. Jee et al. (2007) has shown that combining modalities has important impacts on how people interpret SFUs, which highlights the need to further study how to use utterances within multi-modal interaction.

Similarly to the need to study SFUs as part of multi-modal interaction, it is also important to study how people respond to SFUs when used in different contexts. Real-world social HRI is not context-free. This needs to be highlighted, as it is something that has generally not been accounted for in the prior research on SFUs. From Section 2.3 it is clear that many of the previous works have used an experimental methodology and paradigm within which SFUs were studied without being embedded into a realistic HRI scenario. While on one hand this approach allows a baseline metric/value to be obtained, where the utterances are not biased or confounded by context specific elements, the drawback is that this is not representative of real-world HRI. Thus whether the findings from

[3]Blizzard Challenge has been developed in order to better understand and compare research techniques in building corpus-based speech synthesizers on the same data. (`http://www.synsig.org/index.php/Blizzard_Challenge`)

these experiments would still hold true when SFUs are employed in real-world HRI needs to be investigated further. For example Komatsu (2012) and Read (2014b) have both shown that the context in which SFUs are used impacts their interpretation and the utility that they afford. As such, SFU research can address this drawback by shifting toward conducting experiments and evaluations in paradigms and settings that are representative of what real HRI (in the wild) is like.

Another point is that it is unclear how relationships between people and robots will develop and unfold over time, how robots should be programmed to maintain engaging interactions with people and how software should be designed to exploit the opportunities available to adapt to people and learn from these long-term interactions. With respect to NLUs and GS, one of the main open questions is whether, through prolonged exposure to SFUs, people will learn to make associations between different utterances and their interpreted meanings. In essence, will coherent understanding of utterances emerge from prolonged exposure to utterances (will people proclaim that different utterances have clear semantic meaning associated with them?, i.e. new languages being formed). This is clearly a question that cannot be answered in the near future as robots are not ubiquitous enough for the general population to encounter them on a frequent basis in the real world, and nor does the HRI community understand how to build robots that remain interesting and engaging enough that people readily desire to interact with social robots for pro-longed periods of time and over multiple interaction episodes. Although there are examples, such as Amazon Echo or Google Home, that are having accelerated adoption in the general population with frequent use, these current examples are utilizing Natural Language. Whether the results from the long term interactions with these intelligent personal assistants would hold true with robotic agents using SFUs is still a question that cannot be answered soon.

Classically, research into affective displays through the human voice, as well as synthetic utterances has adopted an evaluation methodology whereby subjects are asked to affectively rate stimuli using explicit methods (e.g. the SAM, Likert Scales, The AffectButton, Categories, etc.) rather than implicit methods (e.g. physiological responses or behavioural changes). As HRI evaluations become more complex as representation of real-world HRI scenarios, it is likely that explicit measurements will no longer be suitable as they demand that subjects attend to both the interaction that they are part of and at the same time providing affective ratings of what the robot is doing. Komatsu (2012) has provided a very good example of this. In order to evaluate whether the NLUs made by their robot were able to convey different degrees of certainty about whether information provided by the robot was accurate, they had their subjects play a game where their performance would depend on how much attention they paid to the utterances made by the

robot. In this case, the quality of the utterances was not measured through subjects explicitly reporting whether utterances conveyed confidence in information, but rather, through the outcome of the game within which the quality of the utterances and the subject's interpretation of these played an important implicit role that impacted the outcome of the game.

2.5 Concluding remarks

In this chapter, research efforts regarding the field of SFU have been reviewed to present the current state-of-the-art and to provide guidelines for future researchers considering to utilize SFU in HRI and especially in social HRI. Reviewing the past research on the field, it has been recognized that despite the commonalities in their objectives regarding social HRI, the auditory interaction modalities other than natural language were not investigated under a single umbrella. To address this need, the notion of SFU has been introduced. The studies utilizing various SFU techniques are grouped under four main categories: Gibberish Speech, Non-Linguistic Utterances, Musical Utterances and Paralinguistic Utterances.

The review clearly shows that although the underlying methodologies have been subjects of psychology, linguistics and musicology sciences for a long time, the SFU research is very young, and especially considering HRI motivated studies, is established mainly in the last decade. Despite the fact that these HRI motivated studies are mostly driven by the application intentions of the robots in question, in many of these studies contextual setting of the HRI is not taken into consideration. Evaluated within or outside the contextual setting, one of the main value propositions of SFUs is being language independent. Even though the cross-lingual and cross-cultural evaluations are rarely performed, the results from these are promising for language independence.

Just as contextual setting, multimodality is an important but understudied aspect of HRI. In essence, multimodality is a natural feature of HRI interaction with gestural, facial, auditory components integrated for affective interactions. The usage of SFUs in multimodal environments is proven to improve the emotion recognition and thus contribute to better affective communication between the human and robot. Naturally the multimodal expression capability of a robot is mainly driven by its morphology. In a number of studies, it has been proven that various morphological characteristics are perceived to have a better fit with certain SFU types and characteristics. So the selection of the best suited SFU type and characteristics for a robot's morphology is an important first step in successful multimodal affective communication. While there are examples of successful use of

SFUs in combination with natural language, the alternative approach of combining multiple SFU types to potentially improve the success of social HRI has not yet been explored.

Child-robot interaction (CRI) as a subcategory for HRI and the use of SFUs in CRI is another promising area of research. Children are more familiar with and more used to affective communication using SFUs, because of its frequent use in cartoons and animation movies, which makes the adaption faster and easier. It is important to recognize that there has not been significant work done to investigate how children interpret SFUs, especially when the contextual setting is taken into consideration.

By introducing the concept of SFUs and bringing together multiple sets of studies in social HRI which have never been analysed jointly before, this chapter of this dissertation addresses the need for a comparative study of the existing literature for SFUs. Considering the short history of this field in the context of social HRI, the already achieved results states a bright future for upcoming research in this space. With this comparative study, multiple promising but currently understudied areas of SFUs have been identified as a guideline for future researchers. The highlighted opportunities for advancements in SFU research may clearly be accelerated further by the parallel progress in social HRI studies in general.

3 Semantic-Free Gibberish Text Generation

3.1 Intoduction

As explained in Section 1.1, this thesis aims at building a framework that allows to study affective human-robot interactions by using vocalizations that do not involve semantics in natural spoken language, namely semantic-free gibberish speech.

The already built acquaintance of people in affective expression through human voice provides an advantage for gibberish speech versus other SFU types, such as NLU or MU. The choice of gibberish speech in this thesis rather than other forms of SFUs is mainly motivated by the assumption that the stronger paralinguistic cues carried by the gibberish speech (Remez et al., 1981) would help in better recognition of the intended emotions during the affective interaction. This is indeed also supported by the results of (Zaga, Vries, Spenkelink, Truong, & Evers, 2016) where they found that children could match the gibberish speech to its intentions with a better recognition rate than the NLU. Also in the same study, it was seen that NLUs led to more ambiguity in the interpretation of the intention, which was also seen in (Read & Belpaeme, 2012).

As a first step towards generating semantic-free affective gibberish speech, a semantic destruction strategy is developed. This allows to create semantic-free texts on which the affective charging capabilities and the naturalness evaluations can be performed.

Several approaches exist to generate gibberish without the semantic content. As Section 2.3.1 lists out, some of those approaches operate on the text transcription and others operate on the speech signal (see Table 2.1). In the HRI domain, the studies have focused mainly on the *cue manipulation via synthesis* approaches operating on the speech signal(Breazeal, 2000, 2002; Oudeyer, 2003; Yilmazyildiz et al., 2006; Yilmazyildiz, 2006; Németh et al., 2011; Chao & Thomaz, 2013). One of the earliest examples of gibberish-like language, was from Breazeal (Breazeal, 2000,

2002) - babbling-like speech for the robot Kismet. She created utterances as strings of various lengths by randomly combining predefined syllable structures, each containing random vowels and consonant. Similarly, Oudeyer's utterance strings (Oudeyer, 2003) were composed of words with various lengths, each containing a combination of open syllables of either CV or CCV type (C = consonant, V = vowel).

Different than the above examples, the approach of constructing semantic-free utterances described in this chapter, starts with an existing intelligible text in a given language, instead of randomly creating it from scratch. This approach has the potential of creating natural language-like gibberish that may evoke the initial language. For effective use in HRI, the generated speech should be high quality, sound lively and not repetitive (Oudeyer, 2003). It can be assumed that, the more natural language-like the resulting semantic-free affective speech is the closer it will be to satisfy being high quality, sounding lively and not being repetitive. At the same time, considering the limited computational resources robots have, the algorithms generating the speech should not be computationally heavy. The approach that will be explained in this chapter aims to serve as a first step to achieve these expectations.

As in many implementations of robots, TTS engines are being used for the auditory output with intelligible speech; the utilization of TTS engines for generating semantic-free utterances was tested in the experiments outlined in this chapter.

3.2 Removing the semantic content from the text

Languages consist of ruled combinations of words and words consist of specially ordered syllables. Syllables are often considered the phonological building blocks of the words and they usually contain an 'onset', a 'nucleus' and a 'coda'. An example of a syllable structure is illustrated in Figure 3.1.

'Nucleus' is usually a vowel-like sound where 'onset' and 'coda' are consonant clusters. The semantic of a word occurs when these vowel nuclei and consonant clusters come together in a certain order by following a set of phonotactic rules of a particular language (e.g. the consonant /q/ is usually followed by the vowel /u/ in English). When the syllabic constituents are modified, the word and consequently the sentence looses its meaning.

An obvious modification would consist of randomly interchanging the vowels and consonants of a word's orthographic representation. However that would generate strange and hard to pronounce combinations. For example, in English

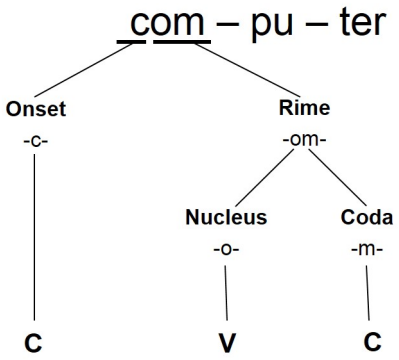


Figure 3.1: Syllable structure for the first syllable of the word "computer".

and Dutch, vowel nuclei are transcribed with one, two and three letters. There are usually only a few vowel graphemes with one letter transcriptions but they are the most frequently used ones in the language (see Figure 3.2). Much more vowel nuclei grapheme sequences exist with two or three letter transcriptions, but these are far more rarely used. A similar relationship also exists for the frequency distribution of the consonant clusters. If the word 'language' would be transformed into gibberish by uniform random substitution of vowel nuclei in the orthography of the word, the result may be something like 'lieungeaugie'. Once the consonants are also swapped, it would likely yield a word like 'sphieudweauthrie' which is a very unusual word (and hard to pronounce).

To avoid such transcriptions, the probabilities of occurrence are calculated for each graphemic sequence of vowel nuclei and consonant clusters in the original natural language, which is English or Dutch in this study.

The probability of occurrence of a vowel nucleus v_i and a consonant cluster c_j in a given text corpus containing n vowel nuclei $V = \{v_1, v_2, ..., v_n\}$ and m consonant clusters $C = \{c_1, c_2, ..., c_m\}$ can be written as:

$$P(v_i) = \frac{f_{v_i}}{\sum_{i=1}^{n} f_{v_i}} \tag{3.1}$$

$$P(c_j) = \frac{f_{c_j}}{\sum_{j=1}^{m} f_{c_j}} \tag{3.2}$$

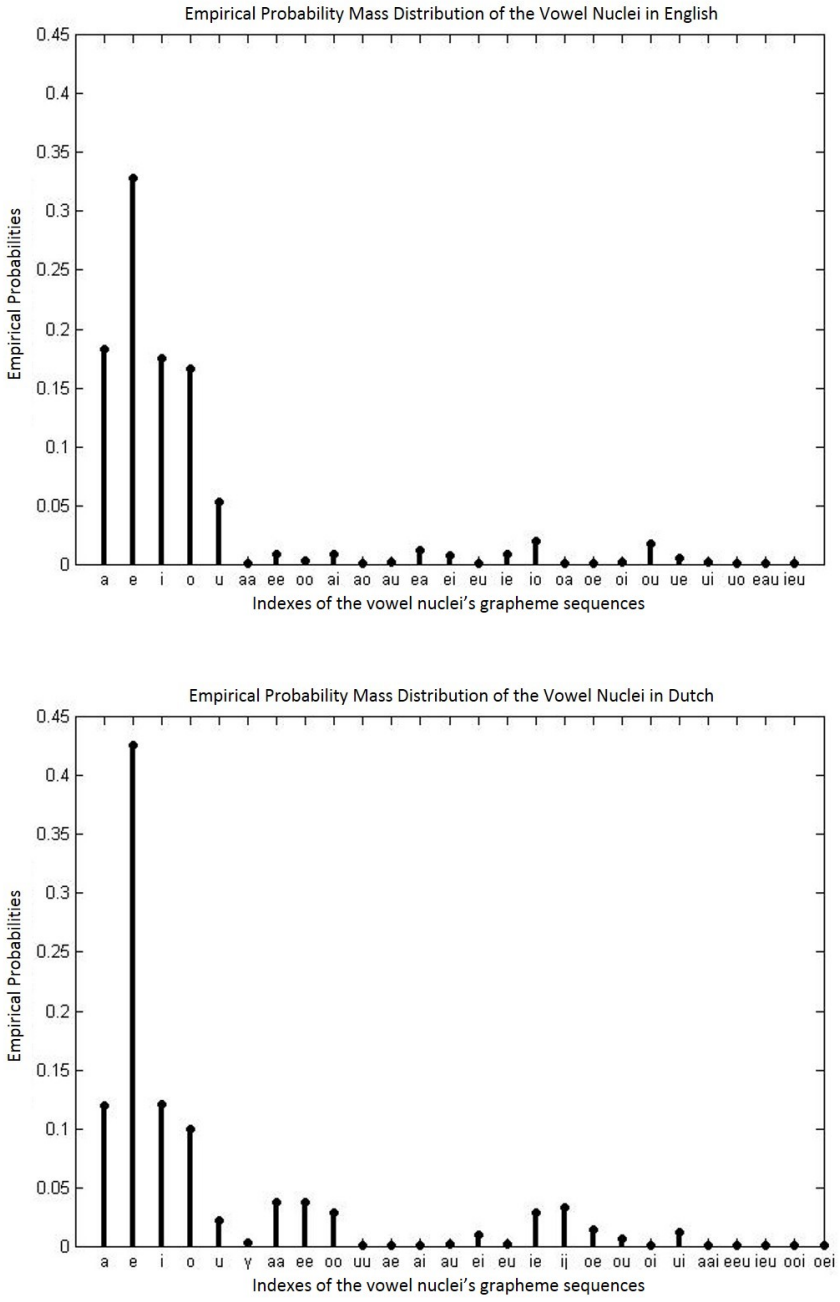


Figure 3.2: Empirical probability mass distribution of vowel nuclei's grapheme sequences in English (upper panel) and in Dutch (lower panel).

in which f_{v_i} and f_{c_j} represent the number of occurrences, or frequency, of the vowel nucleus v_i and consonant cluster c_j, respectively.

For consonant clusters *initial, middle* and *final* consonant cluster probabilities are calculated separately. The probability calculations were performed using English and Dutch texts of approximately 27000 words from a large online text corpus - Project Gutenberg (Hart, 1971).

The graphemic sequences of vowel nuclei and consonant clusters of English and Dutch texts are then replaced by using a weighted selection mechanism in accordance with the calculated natural probability distribution of the graphemic vowel nuclei and the consonant clusters of English and Dutch, respectively.

For example, given the text input to be transformed into semantic-free version is: <This is our beautiful tree>.

- First, the orthographic vowel and consonant letters are marked: CCVC VC VVC CVVVCVCVC CCVV
- Then, the vowel and consonant clusters are identified. VV/VVV/V... type vowel clusters are all considered as one swappable vowel cluster V, and CC/CCC/C... type consonant clusters are all considered as one swappable consonant cluster C, which results in: CVC VC VC CVCVCVC CV
- Next, all the vowel nuclei are substituted with some other vowel nuclei from Figure 3.2 in accordance with the calculated probability distributions. This is done by taking weighted samples from the vector of vowel nuclei probabilities. This turns the input text into: <Th*os es ar bitifal tra*>
- Finally the consonant clusters are also replaced with some other consonant clusters again in accordance with the calculated probability distributions, which finally transforms the input text into: <*Roch ept an siriraf pra*>

In the following two sections, two experiments evaluating the perceived naturalness and emotion conveying capabilities of the resulting semantic-free gibberish are described.

3.3 Perceived naturalness

The approach described in Section 3.2 transforms the existing text into gibberish text. Now one of the questions to be addressed is how *natural* the synthetic speech

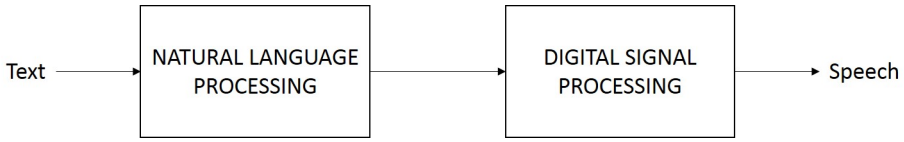


Figure 3.3: A simplified functional diagram of a TTS system.

generated out of these gibberish text strings would sound.

The gibberish text can easily be transformed into synthetic speech with a TTS (Text-to-speech) engine. In general terms, a TTS system can be split-up in two parts as seen in Figure 3.3. . The front-end part contains the Natural Language Processing module and the back-end part is responsible for speech database search and the Digital Signal Processing (DSP).

The Natural Language Processing module takes the raw text as input and outputs a phonetic transcription of the input text together with the desired prosody. This process includes transforming the numbers, abbreviations and acronyms into full text (text normalization); identifying the nouns, verbs, adverbs, etc. (part-of-speech tagging); organizing the text into linguistic units such as clause, phrase, etc. which more closely relates to its expected prosodic realization (syntactic parsing); and constructing the target phoneme sequence for each target utterance (lexicon lookup and letter-to-sound mapping); assigning duration, pitch, accent values and estimating silences between words (prosody generation). The phonetic transcription of the input text and the target prosody values are then used by the Digital Signal Processing module to synthesize the desired physical speech signal.

The Natural Language Processing module of a TTS system is thus language dependent. As they are not designed to work on meaningless text, a set of experiments are performed to investigate:

- How natural the synthetic gibberish would sound
- How the native language of the TTS would affect the result

3.3.1 Stimuli

Two sets of sentences were created as the stimuli. For the first set, the English vowel nuclei probability distributions were used to modify 6 original English sentences. The sentences were selected from children stories of Project Gutenberg (Hart, 1971). For the second set, the Dutch vowel nuclei probability distributions were used to convert 6 sentences that were selected from Dutch children stories of Project Gutenberg. All

Table 3.1: Initial English and Dutch texts and their gibberish versions

English	Gibberish
The king soon married another wife who was very beautiful	Thu kung son merred enithar wofa whi wes vory botufil
In the evening she came to a little cottage	On tho avineng shoa cimo te e lottla cattogo
The rest came running to him and every one cried out	Tha rast cimau rennoung te hom ind iviry ina crad uit
I will let the old lady in	E well lat thi eld ledy en
I need not say how grieved they were	E nid net soy houw grovud thiy waru
The strong way drank the day	Thi string wey drenk tha diy
Dutch	**Gibberish**
Breng het kind naar buiten in het bos want ik wil haar niet meer zien	Brong hat kand nar beten en hot bis wunt iek woel her net mor zan
Zij liep zo lang haar voetjes nog gaan konden tot het bijna avond was	Zo lep ze leng hir vetjas neg gen kendon tet hot beno ovend wes
Zij staken hun zeven lichtjes aan en toen het was het huisje gezellig verlicht	Ze stekaan hen zovaan luchtjes aan on ten het waas hijt hesjo gezeellug verleecht
Je stiefmoeder zal snel weten dat je hier bent	Jaa stefmedar zeel snil wotien det jo her bant
Die was heel mooi om te zien zo met haar rode wangetjes	Di wis hel me em tee zen zi mit har ride wangetjes
Het lief kind was dood en bleef dood	Hat leef kond wos dad een blief doud

12 gibberish sentences (see Table 3.1) were then synthesized both with the English and the Dutch versions of VUB's unit selection TTS (Latacz, Kong, Mattheyses, & Verhelst, 2008). This gave a total of 24 samples categorized in 4 different groups (see Table 3.2): the first group contained 6 gibberish samples created using the Dutch text and synthesized with the Dutch TTS, the second group had 6 gibberish samples created from English text and synthesized with the English TTS, the third and the fourth groups also consisted of 6 gibberish samples each, in which the samples were created with Dutch gibberish text and English TTS, and with English gibberish text and Dutch TTS.

3.3.2 Experimental procedure and participants

Ten subjects (7 male and 3 female) with ages ranging between 24 and 37 participated in a listening experiment. Four subjects had no prior experience with synthetic speech (naive subjects).

The subjects were asked to pay attention to the *naturalness* of the samples.

Table 3.2: The summary of the sample categorization

Sample Group	Initial Text Language	Synthesizer Language
Group 1	Dutch	Dutch
Group 2	English	English
Group 3	Dutch	English
Group 4	English	Dutch

It is assumed that, the more natural language-like the resulting semantic-free affective speech is the closer it will be to achieve effective HRI interaction. Based on this assumption, the subjects were instructed that a sample is to be considered as natural *when it sounds more like an unrecognized real language rather than an unnatural or random combination of sounds.* With this instruction it was intended to guide the users to reflect their opinion on how natural or unnatural the utterances sounded rather than judging the grammar or the content of the sentence. As such the sound of an unrecognized real language in this case is assumed to be a fluent string of speech sounds. They were asked to express their judgment using Mean Opinion Scores (MOS) on a scale of 1 (not natural at all) to 5 (as natural as a real unknown language). They were also requested to identify if the sample sounded like a language they knew.

The samples were provided randomly in a single presentation order. There was no time limit and the participants could replay each sample as much as they wanted. An example of a plain synthetic speech was provided at the beginning of the test to minimize the risk that they would rate the quality of the TTS instead of the naturalness of the gibberish.

3.3.3 Results and discussion

As can be seen in Figure 3.4, the samples created with the Dutch synthesizer had the highest score for both Dutch and English semantic-free gibberish texts. A Wilcoxon Signed Ranks Test indicated that the difference between the scores of the samples created with the Dutch synthesizer and the ones created with the English synthesizer was statistically significant ($Z = -2.416, p < 0.016$). Mean MOS for the samples generated with the Dutch synthesizer was 3.8 while for the samples generated with the English synthesizer it was 3.5. It is likely that subjects were influenced by the synthesizer quality as there is a difference in quality between the Dutch and English versions of the synthesizer. The Dutch synthesis database is almost four times larger than the English database and that difference obviously affects the quality of the synthesized speech. Also almost half of the subjects were native Dutch speakers, which is another possible explanation.

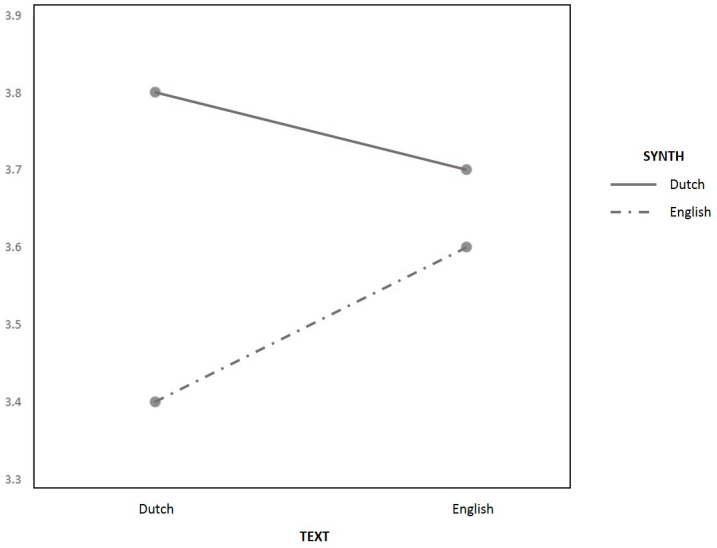


Figure 3.4: Means of the MOS scores on naturalness for all the four synthesizer and initial language combinations. SYNTH = The language used for the synthesis. TEXT = The original language of the input gibberish text

Table 3.3 shows the average MOS scores and Figure 3.5 illustrates the boxplots of the scores for all the 4 synthesizer and initial language combinations. A Friedman test did not indicate a statistically significant difference between these 4 groups ($(\chi^2(3) = 4.778, p = 0.189)$.

Further analysis of the test results, using Mann-Whitney U test statistics showed a significant difference between the overall ratings of the naive subjects and speech processing experts ($Z = -4.511, p < 0.001$). The overall naturalness ratings of the naive subjects were significantly lower than the ratings of the speech experts. According to the feedback given by the naive subjects, it was challenging for them to evaluate the naturalness of gibberish speech independent from the synthesis quality. Although a plain synthesized speech was provided at the beginning of the test to minimize this effect, they indicated that the synthetic speech quality may have negatively influenced their scores.

Regarding the open question of similarity of the samples to a known language, the recognition of the initial language from the synthesized gibberish samples is sum-

Table 3.3: Mean MOS results on naturalness. TEXT = The original language of the input gibberish text. SYNTH = The language used for the synthesis.

SYNTH	TEXT	Mean MOS
Dutch	Dutch	3.8
Dutch	English	3.7
English	Dutch	3.4
English	English	3.6
General Mean		3.6

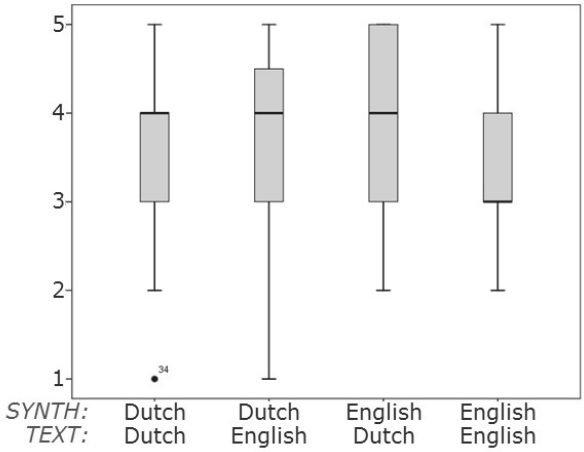


Figure 3.5: Boxplots of the naturalness scores for all 4 experimental groups. SYNTH = The language used for the synthesis. TEXT = The original language of the input gibberish text

marized in Figure 3.6. The recognition rates were highest when both the gibberish input text language and the synthesizer language were the same for both Dutch and English. Again while the text language had a little effect, the synthesizer language had a higher impact on the recognition rates. Dutch was easier to recognize with scores up to 78%. For English the highest recognition rate was about 50%.

A major cause might be driven by phoneme-to-grapheme conversion in Dutch being rather straightforward while the relationship between phoneme and grapheme in English being less regular. For example in English the end of the words "sandwich" and "language" sound the same but they are spelled completely differently. Such a major variation is nonexistent in Dutch language.

Considering the above, the perceived naturalness of the resulting semantic-free gibberish text and the recognition of the initial languages could be further improved by using a swapping mechanism in the phonetic transcriptions, which is closer to the speech than the text. Such a swapping mechanism would first convert the graphemes into phonemes, then would transform the result into semantic-free utterances using phoneme frequencies and then would convert those back to text, or would synthesize the speech directly using the phonetic input if the selected TTS engine would support it. In the context of this framework this topic has been considered as an interesting possibility for future work.

Also the fact that almost half of the subjects were native Dutch speakers and the difference in the quality of the Dutch and English synthesizers, as explained before, could have been other influencing factors for high recognition rates for the Dutch language.

3.4 Influence of semantics on emotion recognition

People naturally use both prosodic meaning and semantic meaning for expressing affect and emotion. In gibberish, there is no semantic information. Furthermore, the fact that gibberish is semantic-free might interfere with the prosodic strategy of the synthesizer and result in less expressive speech. These observations lead to two additional questions to be addressed:

1. How does the semantics of the text influence the perception of the emotions in the synthetic speech?

2. Is gibberish more or less effective than plain speech in evoking the intended emotion?

To investigate these questions, a set of experiments were designed and executed.

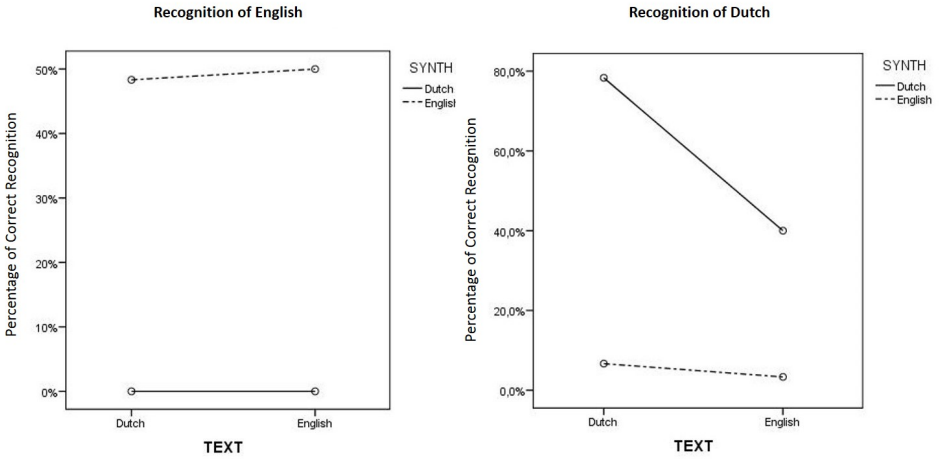


Figure 3.6: Percentages of language recognition. Left panel shows the percentages of the correct recognition of English and the right panel shows the percentages of the correct recognition of Dutch. SYNTH = The language used for the synthesis. TEXT = The original language of the input gibberish text.

3.4.1 Stimuli

Four groups of samples were synthesized for this set of experiments. In the first group, the semantic meanings of the sentences and the acoustic properties of the synthesized utterances corresponded to the same emotion. In the second group, the semantic meanings of the sentences implied the opposite emotion of the acoustic properties. In the third group, the semantic meanings of the sentences were neutral, and in the fourth group the sentences were gibberish and therefore had no semantic meaning. The text input used is shown in Table 3.4.

Two emotion categories were used: *happiness* and *sadness*. 'EmoSpeak', of the synthesizer Mary (Schröder & Trouvain, 2003; Schröder, 2003b), an open source emotional TTS synthesis tool, was used to produce the emotional speech, with the parameter settings for *happiness* and *sadness* reported in (Schröder, Cowie, Douglas-Cowie, Westerdijk, & Gielen, 2001).

3.4.2 Experimental procedure and participants

In this experiment, a forced-choice listening test was performed with nine subjects of age between 26 and 37. The subjects were instructed to listen to a number of samples of which they may or may not understand the meaning and were requested to choose which one of the possible emotions (*happiness, sadness or neutral*) matched the

Table 3.4: Text input for happiness, sadness, neutral and gibberish cases

Happiness	Tomorrow we are going to celebrate my twentieth birthday
	I am so pleased with presents everybody gave me
	My sister's newborn baby is so cute
	After the long winter the sun is finally shining
	I learned this morning that I passed all my exams
Sadness	I didn't do my homework because my dog died yesterday
	It is just a pity that Christmas comes only once a year
	My boyfriend called me to say that is over between us
	This year I didn't get any presents from Santa Claus
	You wish a long goodbye to your friend who is leaving forever
Neutral	There was a picture of a forest hanging in the corridor
	I just read a book about designing listening tests
	I saw a white cat crossing the street
	The man in the restaurant ate a lot of French fries
	In the evening she came to a cottage
Gibberish	Thoru was e pactarai ef e ferest hingang on thi correder
	On tho avineng shoa cimo te e lottla cattogo
	E well lat thi eld ledy en
	Tha rast cimau rennoung te hom ind iviry ina crad uit
	Thu kung son merred enithar wofa whi wes vory botufil

Table 3.5: Confusion matrix for all experimental groups (expressed in %). Rows correspond to the intended emotions and columns correspond to the recognized emotions.

	Inline			Opposite			Neutral			Gibberish		
	Hap	Sad	Neu	Hap	Sad	Neu	Hap	Sad	Neu	Hap	Sad	Neu
Hap	**91**	0	9	**82**	2	16	**76**	0	24	**62**	9	29
Sad	0	**76**	24	2	**56**	42	0	**58**	42	2	**53**	44

sample they heard. The emotive samples were distributed randomly across emotions and a single presentation order was provided to all the subjects. The participants could replay the samples as much as they wanted and there was no time limit.

3.4.3 Results and discussion

Figure 3.7 shows the emotion recognition results for all 4 experimental groups and Table 3.5 summarizes the confusion matrix. Group 1 (semantic meaning and acoustics correspond to the same emotion) had the highest scores among all groups. A Friedman test showed that there was indeed a statistically significant difference in the recognition rates between the different groups, ($\chi^2(3) = 16.123, p = 0.001$). This showed that semantic meaning helps for recognizing the intended emotion, as was expected. However, semantics opposite to the intended emotion did not make the task more difficult than with *neutral* semantics or with *gibberish* speech. At the time of the experiments, the available emotional TTS engines were not yet sophisticated enough to reliably mimic human emotional speech. As such Mary synthesizer simulates *happiness* with high speaking rate and *sadness* with low speaking rate. Therefore the intended emotion could be easily inferred from the speaking rate, which was also provided as a feedback by the subjects. When better performing emotional synthesizers become widely available and easily accessible, these experiments can be performed again to minimize the influence of the synthesizer quality on the results. This is also noted down as part of potential future work.

Importantly, according to a Wilcoxon Signed Ranks Test, there was no significant difference between the samples with an emotionally neutral meaning and the gibberish samples ($Z = -1.333, p = 0.182$), showing that the emotions were conveyed with gibberish speech with a similar performance as with semantically neutral speech.

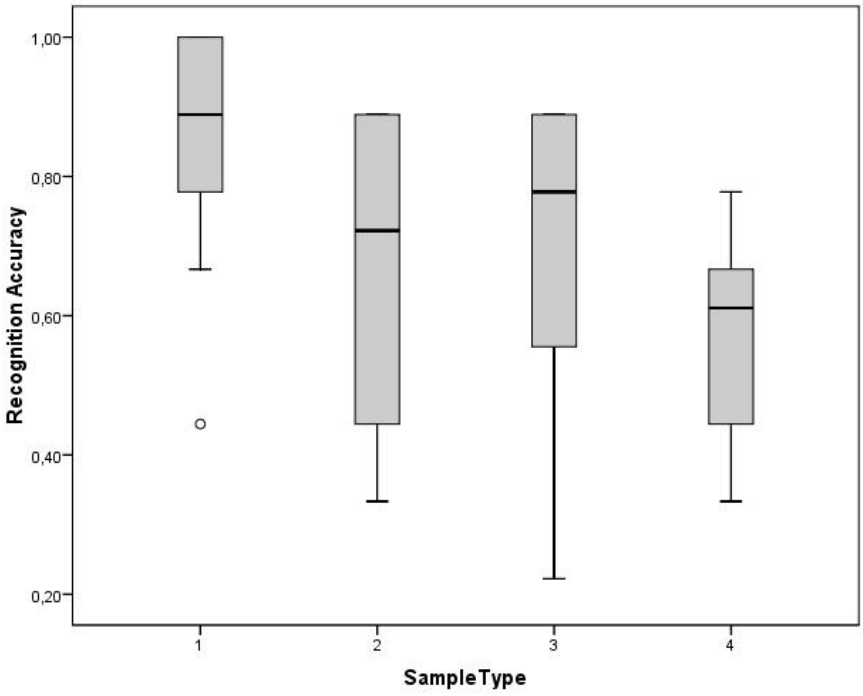


Figure 3.7: Box plot of the emotion recognition results for 4 different experimental groups. (1 = Semantic meaning and the acoustic properties of the utterances correspond to the same emotion, 2 = Semantic meaning and the acoustic properties of the utterances have opposite emotions. 3 = Semantic meaning of the utterances are neutral, 4 = The utterances are gibberish and therefore have no semantic meaning)

3.5 Summary

This chapter has described the approach that allows to create semantic-free gibberish text as a component of the Semantic-Free Affective Speech Framework and presented the results of two experiments evaluating the naturalness and the affective charging capabilities of the resulting gibberish.

The semantic of an existing text in a language is destroyed by replacing the vowel nuclei and consonant clusters of the text using a weighted selection mechanism in accordance with their natural distribution in the same language.

In the first experiment, the influence of input text and language on the synthesized gibberish speech was explored. It was seen that the gibberish speech created resembles a natural language with a total average MOS of 3.6 out of 5. That is important since the goal is to create a semantic-free speech that sounds like a real language and not as an unnatural or random combination of sounds.

At the same time, some subjects reported that the synthesizer quality might have affected their scores. This was also evident in the results as the samples generated with the Dutch synthesizer, which had a higher quality, had higher scores than the samples from the English synthesizer.

The experiments also showed that the semantic-free gibberish speech resembles the source language when good quality synthesis is used in combination with an input text from the same language. This can be easily understood considering, the synthesizer still uses the phones and intonation model of its target language, even when synthesizing text without semantic meaning.

From the second set of experiments on the relation between semantic meaning and the perceived emotion, it was found that semantics help recognizing the intended emotions when the semantic and the prosodic meaning of the utterances are both in line with the intended emotion. When they were in line with opposite emotions, this did confuse the subjects but less so than might have been expected. A probable cause would be that it was quite clear that the synthesizer used simulates *happiness* with a faster speech rate and *sadness* with a slower speech rate so that the emotion setting of the synthesized utterance could be easily detected. There was indeed feedback from the subjects that they mostly rated according to the speech rate.

When it comes to semantically neutral samples, no statistical difference was found between samples with emotionally neutral meaning and gibberish samples. Thus, when there is no emotional meaning in the text, it did not make much

difference for the recognition of the intended emotion whether a meaningful or a gibberish utterance was used to communicate the emotion.

Some of the techniques, experiments and results mentioned in this chapter have been published in (Yilmazyildiz et al., 2010; Yilmazyildiz, Verhelst, & Sahli, 2015).

4 Semantic-Free Affective Gibberish Speech

4.1 Introduction

The experimental evaluations in the previous chapter showed that gibberish speech resembles a natural language and that it is as effective as semantically neutral speech in communicating the intended emotions. However, in both of the tests, subjects reported that the synthesis engine quality affected their evaluations. This highlights some of the problems of using TTS systems to synthesize emotions:

- The final expressive speech strongly depends on the TTS engine quality and the quality of the expressivity models of TTS engines at the time were not yet mature enough.

- The *voice quality* is an important factor for communication of emotions (Murray & Arnott, 1993; Schröder, 2001). However, the voice quality of the emotions is not fully transmitted to the synthesized speech using the currently available TTS engines.

To overcome these drawbacks to a large extent in the SFAS framework, a data-driven method has been developed. The aim of this data-driven method is to synthesize high quality semantic-free affective speech utilizing a limited duration of recorded gibberish speech.

As can be seen in Figure 4.1, an emotional database is at the core of this method. The aimed semantic-free affective speech can be achieved either by playing back the utterances in the database or by synthesizing more unique utterances with segment swapping, which is a concatenative synthesis technique as will be described in Section 5.2. For both of these options an affective speech database which is already in semantic-free form is required. This database is aimed to achieve acceptable levels of emotion recognition by recording affective gibberish speech that naturally incorporates the voice quality of the emotions, which also bypasses the shortcomings of the dependence to the TTS engine quality and the quality of the

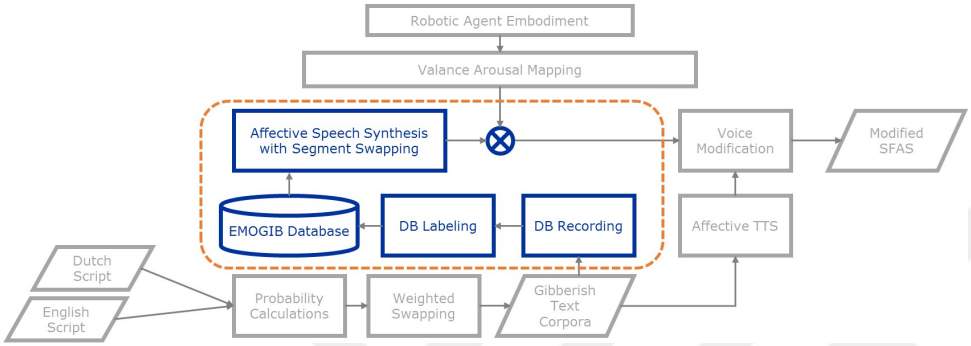


Figure 4.1: Data-driven method (highlighted with the dashed line) in the framework

expressivity models of the TTS engines. For an effective use, the meta-data information, such as emotion labels and various prosodic segment units in different lengths is also identified in a database labeling step which will be described in Section 5.2.1.

This chapter focuses on the emotional gibberish speech database. The various steps needed to construct this new emotional gibberish speech database will be described and the assessment of the quality of the emotions as well as the naturalness of the utterances will be presented in the following sections of this chapter.

4.2 Brief outlook on the existing expressive corpora

There has been a considerable amount of work in recent years, on the collection of auditory, visual and audiovisual emotional data. Researchers in this space have been constructing expressive databases that answer specific research questions. The focus on answering specific research questions leads to challenges in reusing existing databases for research that focuses on other sets of questions or challenges.

The majority of the available emotional speech databases in the literature are for emotion recognition purposes. They are usually multi-speaker databases which contain many samples per emotion but few samples per speaker (Batliner et al., 2004; Breazeal & Aryananda, 2002; Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005; Douglas-Cowie et al., 2007; Grimm, Kroschel, & Narayanan, 2008; Liberman, Davis, Grossman, Martey, & Bell, 2002). Additionally, some of them contain naturally occurring spontaneous emotions that are collected from television shows, telephone conversations, etc., where the qualities are not suitable for emotion

synthesis research (Batliner, Hacker, Steidl, Nöth, & Haas, 2003; Douglas-Cowie, Cowie, & Schröder, 2000; Hansen, Bou-Ghazale, Sarikaya, & Pellom, 1997). There are larger good quality emotional speech databases that are available for synthesis purposes (Ambrus, 2000; Iida & Campbell, 2003; Iriondo, Planet, Socoró, & Alías, 2007; Saratxaga, Navas, Hernáez, & Luengo, 2006), however those ones are not semantic-free. Few exceptions are GVESS - Geneva Vocal Emotion Expression Stimulus Set (Banse & Scherer, 1996) and GEMEP-Geneva Multimodal Emotion Portrayals (Bänziger & Scherer, 2010; Bänziger, Mortillaro, & Scherer, 2012) corpora which contain gibberish utterances. However in these datasets the same text script was used for each emotion as the purpose was to obtain acoustic profiles of vocal parameters for different emotions and that makes it unsuitable for the scope of this research.

Thus, a new gibberish emotional speech database suited for high quality expressive semantic-free gibberish speech synthesis was recorded.

Readers are pointed to the extensive reviews of the available expressive corpora from (Cowie, Douglas-Cowie, & Cox, 2005; Douglas-Cowie, Campbell, Cowie, & Roach, 2003; Ververidis & Kotropoulos, 2003).

4.3 EMOGIB: Emotional gibberish speech database

EMOGIB is an expressive gibberish speech database that contains approximately 15 minutes of speech (∼1800 words) for each of the big six emotions (*anger, disgust, fear, happiness, sadness, surprise*) and 25 minutes of speech (∼4100 words) for *neutral* state. It has 4 different semantic-free gibberish corpora: C1 & C3 - generated by using the whole consonant and vowel space of Dutch and English, C2 & C4 - generated by using the whole vowel space and voiceless consonant space of Dutch and English. The reason of generating C2 & C4 comes from the ease of using voiceless consonants for automatic labeling and manipulation.

The requirements considered in the design of the EMOGIB database are briefly described below:

- Controlled variation in the text scripts: To be able to reflect the varying affective charging capabilities of short and long sentences, the corpus should include sentences containing different number of words. Also in each emotion category the proportion of the number of words should be similar.

- Voice type suitability: Considering the aimed primary usage of the resulting database, the voice type should suit a robotic character.

- Voice type stability: The voice type should be kept stable during the entire recording session.
- Stability of the emotion quality: The voice quality of the emotion should be consistent across the entire recording for each emotion.
- Limited complexity for gibberish framework validation: Not to over-complexify the framework and the validation process, the initial system should be limited to the widely used basic emotions (*anger, disgust, fear, happiness, sadness, surprised*) and *neutral*.

Encompassing all of the above, the goal was to create a large expressive database that is in semantic-free gibberish form and sounds like a real language. To achieve this goal, the content of the corpus and the speaker were carefully selected and special attention was paid in the recording procedure to have a good quality across all the emotions in scope.

4.3.1 Speaker selection

Many of the requirements that effect the quality of the final database are influenced by the acting qualities of the selected speaker. Even though it is possible to improve the performance of the speaker by carefully designing the recording conditions (Busso & Narayanan, 2008), the speaker selection is still a key factor and requires attention.

A call for speakers was distributed to the theater/drama schools in the country. Six candidates were invited to a phone interview. The candidates were all informed before the interview that they would be asked to voice-act in the interview. Four sentences were sent to them (one English, one Dutch and two gibberish sentences) that might be used as scripts to voice-act.

Each interview started with a friendly talk where their personal information such as their name, age, study program, languages spoken, experience in voice acting was gathered. The questions in the second part of the interviews were structured in a way that the candidates could be evaluated on the following criteria: the ability to easily switch the voice to another type, the ability to act emotions, the ability to act gibberish sentences, the flexibility of the voice, the capability of maintaining the voice quality during the recording session and the ability to act as fitting certain characteristics of an imaginary robot (such as *humor, pleasure, funny, stupid, emotional, sympathetic*). This was to evaluate their ability to easily switch the voice to another type as well as their ability to act as fitting the required characteristics. To judge their ability to act emotions and their ability to act semantic-free sentences,

they were instructed to act the scripts that were sent prior to the interviews in six basic emotions (*anger, disgust, fear, happiness, sadness and surprise*). Finally, to assess the flexibility/limits of their voice, they were requested to act certain ages and genders such as *male, female, child, old man, old lady*. All the interview sessions were conducted through an Alcatel-Lucent 4019 phone in hands-free mode and the sessions were recorded to be able to listen to them later for evaluation.

Based on the above criteria, a 20 year old female drama student was selected as the speaker for the actual recordings.

4.3.2 Text corpus

Four corpus sets were created for the recordings, each set containing 7 different script sets (one for each emotion category and one for the neutral category). The first corpus set was generated by replacing the entire vowel nuclei and consonant clusters in the selected Dutch texts using the weighted selection mechanism detailed in Chapter 3 in accordance with the natural probability distribution of the *vowel nuclei* and the *consonant clusters* of Dutch. For generating the second corpus set, the entire consonant clusters in a Dutch text were replaced in accordance with the natural probability distribution of *voiceless consonant clusters* of Dutch while the vowel nuclei were replaced in accordance with the natural probability distribution of the *vowel nuclei* of Dutch. The third and the fourth corpora were created accordingly but this time using English texts and the corresponding probability distributions of *vowel nuclei*, *consonant clusters* and *voiceless consonant clusters* of English. The structure of the four corpus sets are summarized in Table 4.1.

The probabilities of occurrence in English and Dutch were calculated (as described in Section 3.2) for each vowel nucleus and for each consonant cluster. For consonant clusters begin (onset), middle and end (coda) consonant cluster probabilities were calculated separately. Similarly, the same calculations were performed for the voiceless consonant clusters (begin, middle, end).

Table 4.1: The summary of the corpora structures

Name	Language	Consonant Distribution	Vowel Distribution
C1	Dutch	Whole consonant space	Whole vowel space
C2	Dutch	Voiceless consonant space	Whole vowel space
C3	English	Whole consonant space	Whole vowel space
C4	English	Voiceless consonant space	Whole vowel space

The texts were categorized in a way that there would be controlled variation in the sentences. These sentences contained different number of words, starting from one word up to ten words. In each emotion category the proportion of the number of words was the same. The sentences were organized in paragraph structure to provide a dialogue impression (see Table 4.2).

Table 4.2: An example of script paragraph structure from corpus C1 to provide dialogue impression, for anger in semantic-free form

- Gocht en bij heep! Deang a h dup. Men iet ge wijechtilutsprajcht! En gaar doet he? O dons be arkanvae. Deevraen en oingelete stiar. Rer lo vaaroe beraan! Deen staar vean gee. Kan ing deit. Goors e den eeds?
- Hoeen e gost oos an ve mijm doaj, wahael bienfey? Ga denhijul tieg allieder en nian blooot sned ien eengee rop. Irieds o dej sprorbecten zens vian ui d buronkees dej! We bend ist e biaf twue noopo feon vemst geetut
...

4.3.3 Database building

4.3.3.1 Recording setup

The recordings took place in a professional audiovisual studio located at the university campus (*ETRO Audio-Visual Lab*, n.d.) where the proper acoustic absorption was provided. The speaker was sitting on a stool chair with proper headphones. The microphone (Neumann U87) was at a fixed position from the mouth of the speaker. Reading pane was put at a position where the speaker felt comfortable. Figure 4.2 shows the recording set up.

The control room (Figure 4.3) where the monitoring of the recorded signals and the controlling of the prompter were done was outside the recording chamber and there was a window connecting the rooms visually.

4.3.3.2 Recording procedure

The recordings started with voice tuning practices. The voice type should have suited the robotic character. On the other hand, as the speaker would use the same type of the voice for a long period of time, it was important to find the voice type that the speaker felt comfortable with. Prior to the actual recordings, the speaker improvised a few different voice types and they were all recorded. Considering the above two criteria, one of the voice types was chosen as the base voice in consultation with the speaker. During the recordings, the recorded sample of the

Figure 4.2: Overview of the recording setup

Figure 4.3: The control room where the monitoring of the recorded signals and the controlling of the prompter were done

voice type was periodically played back to the speaker, in order to keep the voice type stable during the entire recording session.

The same reference building procedure was repeated before each emotion recording. Taking the recorded base voice as a reference, the speaker improvised each emotion with that voice type. Then the final sample was kept as a reference for that emotion and the actress practiced for a while. At the beginning of each script paragraph, the reference was played and the speaker continued acting in the same voice quality of the emotion. Also during the recordings, whenever a difference in the level/quality of emotion or voice type was noticed, that part was compared with the reference and re-recorded if needed.

A stuffed prototype of the robot Probo (one of the evaluation platforms of the SFAS framework) was put in the recording room. This helped the speaker to act as being a robot. The photographic facial expressions of the robot were pinned on the face of the stuffed prototype to visualize the robot's emotions. The speaker found that method helpful for staying in the mood of the intended emotion.

Before the recordings, a short discussion was held with the speaker about how to get in the mood for the different emotions. The speaker was also a drama trainer for children. She shared that in their acting trainings, they let the trainees close their eyes and recall some scenes from their lives that had the particular moods/emotions. The same method was used to get herself into the mood. Only when she could not bring any scene from her life, a short story in that particular emotion about Probo was shared with the actress.

The speaker chose the emotion as well as the text corpus to start with. 5-10 minutes of breaks hourly were planned but the speaker could also take a break whenever she felt the need.

The recordings were done with Pro-Tools 8 and the pre-amplifier used was Earthworks 1021. All the data is recorded with 48 kHz sampling rate and 24 bit.

4.4 Evaluations

4.4.1 Experimental procedure and participants

A series of two experiments were performed; one with adult listeners and one with children.

While more subjects participated in the children experiment (thirty-five subjects

with ages ranging between 10 and 14), the audio part of the children experiment was structured as a subset of the adults experiment. Only one database subset (C1) was used in the children experiment considering the shorter attention span of the children. Aside from the audio section, the children experiment also included visual and audiovisual sections. Because of this, the children experiment is analyzed and discussed in more details in Section 6.3 while this section focuses primarily on the adult experiment.

Ten subjects participated in the adult experiment. The age of the subjects varied between 27 and 32.

It would have been ideal to evaluate all the samples in the database. However due to the large size of the database, random samples were selected from each database subset (C1, C2, C3, C4) for each emotion category. The length of the samples had to be long enough so that the subjects could evaluate effectively. On the other hand, the length should not be too long not to lose the attention of the participants. Thus, four samples of 10 seconds were selected to be used for each emotion.

The subjects were instructed to listen to a number of samples of which they might not understand the meaning. The order of the samples were distributed randomly across emotions and a single presentation order was used for all the subjects. The subjects were requested to choose which one of the possible emotions *anger*, *disgust*, *fear*, *happiness*, *sadness*, *surprise* or *neutral* matched the speech sample they heard. Subjects could listen to the samples as many times as they needed.

As the final goal is to create a *natural sounding* semantic-free gibberish speech that can be used in building expressively interacting computing devices, also the naturalness of the database had to be evaluated. Thus, in a second question, the subjects were asked to pay attention to the naturalness of the samples. They were instructed that the sample was considered as natural *when it sounded rather like an unrecognized real language and not as an unnatural or random combination of sounds*. With this instruction it was intended to guide the users to reflect their opinion on how natural or unnatural the utterances sounded rather than judging the grammar or the content of the sentence. As such the sound of an unrecognized real language in this case is assumed to be a fluent string of speech sounds. Subjects were asked to assess their perception of the naturalness of the samples using Mean Opinion Scores (MOS) in a scale from 1 to 5. They were also asked to write down the language if the sample sounded like a language they knew. That was to investigate if it was still possible to recognize the original language of the corpora after consonant and vowel swapping.

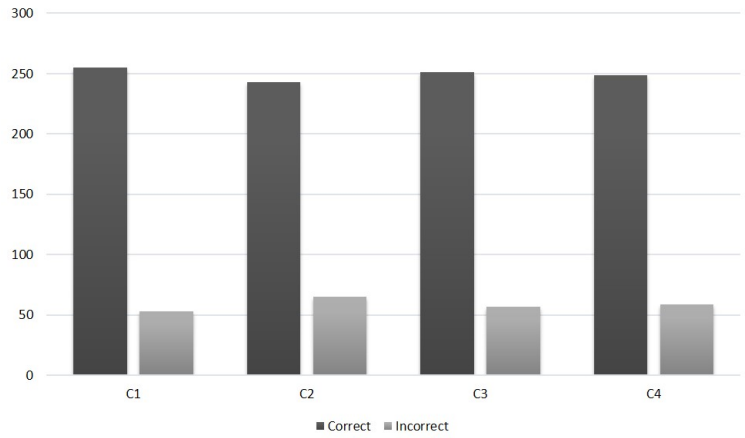


Figure 4.4: Emotion recognition results for all 4 experimental corpora. x-axis corresponds to correct-incorrect emotion recognition grouped by each corpus and y-axis corresponds to the number of samples.

4.4.2 Results

Figure 4.4 shows the emotion recognition results for all 4 experimental corpora (C1, C2, C3, C4). "Correct" stands for the emotion that was perceived as the intended emotion and "incorrect" stands for the emotion that was perceived as one of the other emotions and not the intended one. As can be seen from the graphs, there is not a big difference in the recognition results which was also confirmed by a Friedman test ($\chi^2(3) = 1.648, p = 0.648$).

Overall intended emotions versus recognized emotions are shown in the confusion matrix of Table 4.3. *Sadness* was recognized by most of the participants (94%). The recognition rate of *sadness* was followed by *neutral* with 88%, *surprise* with 87%, *happiness* with 84%, *disgust* with 74%, *fear* with 73% and *anger* with 66%. *Fear* was usually confused with surprise, and anger with *neutral* or *surprise*.

In the children experiment, in which only C1 was used, *sadness* was recognized best (100%). This was followed by *surprise* with 86%, *fear* with 71% and *disgust* with 57%. *Happiness* was often confused with *anger* and vice-versa which resulted in a lower recognition (29% and 46%, respectively). Much better results were achieved in the adult experiment for the same samples of corpus C1 (91% and 64% for *happiness* and *anger*, respectively). This difference can be an indication that children and adults might have a different interpretation of, especially, *happiness*. And further research is needed to check this hypothesis. For the other emotions,

Table 4.3: Overall confusion matrix of the experiment with adult subjects (expressed in %). Rows correspond to the intended emotions and the columns correspond to the recognized emotions.

	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Neutral	**88**	1	0	0	10	1	1
Anger	13	**66**	3	1	6	1	10
Disgust	5	8	**75**	1	2	8	2
Fear	0	6	0	**73**	1	7	12
Happiness	2	1	1	3	**84**	6	5
Sadness	1	1	0	4	1	**94**	0
Surprise	2	3	1	2	1	4	**87**

Table 4.4: Experimental results on naturalness of the experiment with adult subjects

Corpus	Mean MOS
C1	3.5
C2	3.3
C3	3.4
C4	3.3
General Mean	**3.4**

the recognition rates for C1 in the adult experiment were: 100% for *sadness* and *surprise*, 91% for *fear*, and 55% for *disgust*.

Table 4.4 shows the average MOS scores and Figure 4.5 illustrates the boxplots on naturalness for each corpus. As can be seen, the overall mean score is 3.4. The MOS result of corpus C1 was slightly higher than the other corpora but the difference was statistically significant only for C2 ($Z = -2.836, p = 0.005$) and C4 ($Z = -2.824, p = 0.005$) based on the Wilcoxon Signed Ranks Test using Bonferroni correction [1].

Figure 4.6 shows to what extent the subjects were able to identify the original language in C1 and C3, where the natural distribution of both vowels and consonants were used. It was seen that, for most of the subjects both of the corpora did not sound as any language they knew. For the samples that the subjects thought they had recognized an existing language, the majority of them suspected these to be Dutch or English, for C1 and C3 respectively.

[1] Bonferroni correction is calculated by dividing the significance level initially being used by the number of tests that were run to gather the new significance level. In this case the new significance level is: 0.05/6= 0.008

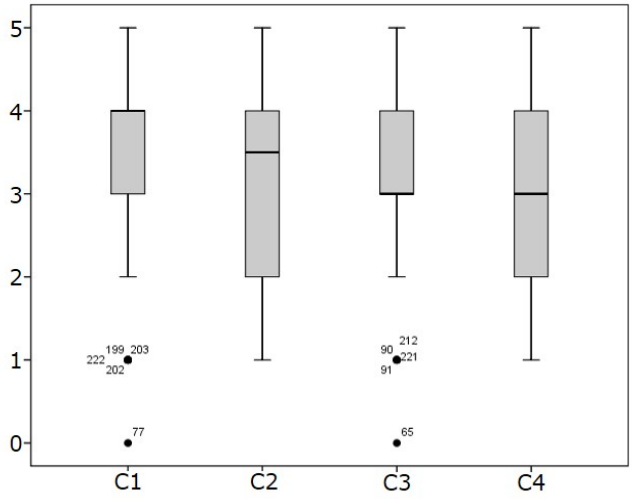


Figure 4.5: Box plot summarizing the naturalness scores for each corpus.

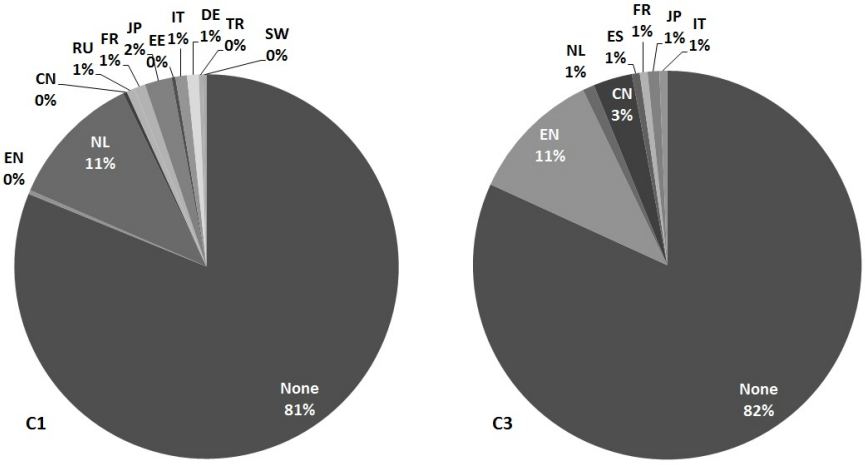


Figure 4.6: Percentages of language recognition for C1 and C3 corpora.

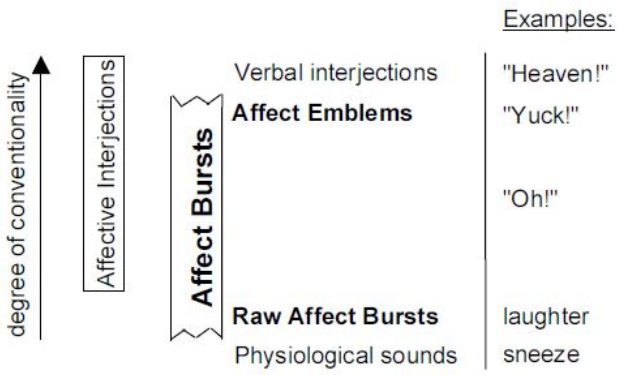


Figure 4.7: Categorization of affect bursts and interjections (Schröder, 2003a)

4.5 Extensions on EMOGIB

In expressing emotions, humans naturally use various short, affectively charged, non-speech vocal sounds and expressions with some degree of phonemic structure in addition to the speech. These sounds are referred to as affective bursts (K. R. Scherer, 1994). For example "Yuk!" for disgust, "hiii?" for surprise, or "rrrrrr" for anger, etc. There also sounds, which might be referred to as verbal 'interjections' or 'fillers', that are short phonemic structures, heavily charged affectively. These interjections are commonly used in combination with affective bursts and the borders between the two expression categories can be blurry (Schröder, 2003a). The categorization is visually represented in Figure 4.7.

Inclusion of these sounds into the EMOGIB database, may potentially increase the naturalness of the speech and the perception of the emotions from the Semantic-Free Speech.

Another extension towards further improved naturalness would be a natural dialogue impression. A natural dialogue impression can be created by using a structure similar to dialogues used in theater/film scripts which the actors are familiar with, so that the impression of a monologue would be minimized. This structure already existed in the first set of EMOGIB database to some degree. The sentences were presented to the speaker in paragraph structure to provide a dialogue impression. But more intensive dialogue structure would be desirable.

Because of the two main needs explained above, an extension on EMOGIB was designed and new recording sessions were performed with the same speaker by

repeating the same recording procedure as outlined in Section 4.3.3.2. New scripts were created for the recordings. The scripts were taken from original Dutch/English plays which were already in dialogue form and which included the places of the paralinguistic sounds on the scripts in brackets. Those original Dutch and English scripts were then transformed into semantic-free gibberish. The extension provided additional semantic-free speech of approximately 3 minutes per emotion. However, evaluation experiments need to be performed before further integrating this affective interjection enriched data set into the EMOGIB database, which is considered as future work.

Table 4.5: An example of theater script structure from C3 to further improve dialogue impression for anger in semantic-free form

SWIONAND — O prest ar mos wio pastieri!
ION — Iadumbly ir rarebly di at valy thu tong theuf theeging.
(man o cioncoiwry thaucta) Ee doffir, thuinents oudy an r teffi jif e cheory helt etung thou. Heu mest'h wertevu outty.
SWIONAND — ...

4.6 Summary and discussion

This chapter described the EMOGIB emotional gibberish speech database with its primary aim of affective communication between robots and their users including children as part of the broader SFAS framework, along with the evaluations on the recognition of the emotions.

4.6.1 Database design

Special attention has been paid in building the database. Douglas-Cowie et al. defined four main areas that needed to be considered in the design of such a database: scope (number of speakers, emotional classes, language, etc), naturalness (acted versus spontaneous), context (in-isolation versus in-context) and descriptors (linguistic and emotional description) (Douglas-Cowie et al., 2003). In this section, the EMOGIB database is discussed and summarized in terms of these areas.

Scope: As the EMOGIB database was designed for affective speech synthesis, recording only with one actor was suitable to record the required expressions, as also reported in (Douglas-Cowie et al., 2003). Regarding the emotions covered, the widely used basic emotions (*anger, disgust, fear, happiness, sadness, surprised*)

and *neutral* were chosen to be utilized in the initial system. The database could be extended for other emotions in case required.

Naturalness: As mentioned by Douglas-Cowie et al., the price of control over the data is the naturalness (Douglas-Cowie et al., 2003). It should be noted that different from the previous uses in this dissertation, the term "naturalness" in this current evaluation area corresponds to acted versus spontaneous attributes of the emotions and not to sounding like a real language versus random combination of sounds. In the EMOGIB corpus, the tradeoff between the control over the data and the naturalness was attempted to be balanced by selecting appropriate material designed as paragraph structure in the first edition and as dialogue scripts in the extension version which were requested from the speaker to be acted as in a play. With these settings, some natural realizations of emotions that are not observed either in monologues or in read speech material can be observed. In this sense, this database might be labelled as semi-natural (Douglas-Cowie et al., 2003) as an actress was used for the recordings, who might have exaggerated the expression of the emotions but based on the setting used to elicit emotions, the emotional quality might include spontaneous emotions as well.

Context: The fact that the text scripts were gibberish, eliminated the benefit of semantic context (such as the tendency of vocal cues to follow emotionally significant words) that might be available in spontaneous speech easing decoding the emotions. However, gibberish has the advantage of allowing to fill in the blanks of the semantics. As also reported by the speaker, it is possible to imagine any semantic content on gibberish text scripts. In terms of the structural context, various characteristic of the utterances (long or short phrases) exist in the text scripts of the database, allowing to capture variations in emotional tone as suggested by (Douglas-Cowie et al., 2003).

Descriptors: Anger, disgust, fear, happiness, sadness, surprise and *neutral* are basic emotion descriptors also used in the evaluation of the selected samples of the database by the listeners. The database is segmented at the sentence level and, in terms of linguistic descriptors, the graphemic transcriptions of the segmented utterances are available. Also, as part of the data labeling step; the voiced, unvoiced and pause speech segment boundaries for each utterance are identified automatically which will be described in more detail in Section 5.2.1.

In summary, the EMOGIB speech database was designed to satisfy the key requirements presented in Section 4.3. As a result, this database contains natural-like semantic-free gibberish speech that can be used in various affective communication applications as part of the SFAS framework.

4.6.2 Evaluations

The perception experiments showed high emotion recognition results of up to 81% overall (and even up to 94% for certain emotions) which were better results than seen earlier in the field (Breazeal, 2000; Oudeyer, 2003; Tickle, 2000).

Across the four unique corpus sets, no statistically significant differences were found in the overall recognition results. This means that the applied methodology of recording induced emotions of an actor helped achieving stable recognition results. The main driving reason for this stability can mostly be attributed to the utilization of the control/reference sentence which was described in Section 4.3.3.2.

This high decoding accuracy, in comparison to earlier results in the field, strengthens the argument on the importance of the voice quality and the high-quality database in affect expression.

It is seen that the semantic-free gibberish speech created resembles a natural language with an overall mean score of 3.4 on a scale of 1 to 5. That is important since the goal was to create semantic-free speech that sounded like a real language. Unlike the emotion recognition, the score cannot be compared easily with the scores from the literature. Naturalness as a concept in the literature can be considered underspecified (Dall, Yamagishi, & King, 2014). There is not an exact aligned definition of what naturalness is. As differing studies give participants differing instructions, comparison between the naturalness results from various studies or assuming a certain baseline is not feasible. In this study, a possible baseline could be the ratings for a real unknown foreign language. The results of the naturalness scores from this study could then be interpreted more accurately once such a baseline is available in the literature.

In general, the gibberish speech created does not sound as any other language known by the subjects. For the corpora where a natural distribution of consonants and vowels was used (C1 and C3), the gibberish speech still sounded slightly like the languages of the texts that were used to create the gibberish texts.

No statistically significant differences were found between the four different corpora for emotion recognition results. For the naturalness, C1 was better performing with a small margin (+/- 0.2 in MOS) compared to C2 and C4. Considering all four corpora had a MOS of 3.3+ (out of 5) for *naturalness*, they can all be utilized for emotional speech communication studies. Combining the results from the adult experiment that the gibberish speech resembled a natural language with an average MOS of 3.4 (out of 5) with the results of the children experiment that they liked

the voice with an average MOS of 7.0 (out of 10), this database can be used in further studies across subjects with various age groups.

Some of the techniques, experiments and results mentioned in this chapter have been published in (Yilmazyildiz et al., 2011, 2015).

5 Speech Modifications

5.1 Introduction

The previous chapter has described the construction of Semantic-Free Affective Speech (SFAS) as a core component of the SFAS framework that allows to study affective human-robot interactions. Also the design and building of the EMOGIB database that contains emotions with this new speech have been detailed. The new semantic-free gibberish speech was shown to resemble natural language.

This chapter focuses on two speech modification techniques that are instrumental to further advance the framework in social HRI studies: segment swapping and voice modification.

The segment swapping section explains the concatenative synthesis mechanisms to expand the number of unique semantic-free utterances and evaluates whether these modifications would harm the emotion perception and the naturalness of the resulting speech samples. The voice modification section details the algorithms that provide the alignment of the voice characteristics of the SFAS with the robot morphology.

As mentioned before, EMOGIB as an expressive gibberish speech database contains approximately 15 minutes of speech for each of the big six emotions (*anger, disgust, fear, happiness, sadness, surprise*) and 25 minutes of speech for *neutral* state. Building such a database requires a significant effort, especially during the recording phase. While the segment swapping and voice modification techniques that are detailed in this chapter primarily focus on further advancing the SFAS framework, the segment swapping can also provide a guidance on the optimal recorded speech duration for researchers who would utilize the framework in generating their own expressive semantic-free speech databases.

These expressive semantic-free speech databases are mostly intended to be utilized through robotic agents. Like humans, robotic agents also have various physical attributes that complement their physical morphologies. The effects of any misalignments between these attributes and the robot's physical morphology on

observers' acceptance of the robotic agent is a fairly well studied issue in the HRI field. However voice style, as one of those attributes that require alignment with the physical morphology of the robot, hasn't taken as much attention as other physical appearance related attributes. This might be partially due to Mori's widely known but not fully accepted Uncanny Valley(Mori, 1970) hypothesis focusing mainly on the physical appearance of the robot. There are a few notable studies focusing on the vocal attributes. (Mitchell et al., 2011) has explored the human realism of a character's visual elements and its synthetic or human voice, concluding that they should match. In (Read & Belpaeme, 2010) observers felt more comfortable with human like voices for humanoid robots. (Komatsu & Yamada, 2011)'s study expressed that the robotic agents' appearances may even affect people's interpretations of the agents' expressions, even though these agents express the same information. What hasn't been explored so far and is the aim of the voice modification section, is finding the matching voice style (specifically voice spectral shift) for a robotic agent in alignment with the physical morphology of the robot.

5.2 Segment swapping

The EMOGIB database that was described in detail in Chapter 4 was constructed as an important component of the framework. It is aimed to be a natural sounding non-semantic vocal communication medium for robotic agents to convey (simulated) emotions. However the total duration of the unique semantic-free utterances from the EMOGIB database is constrained by the total duration of the recorded speech. Based on the initial experiments, the 15 minutes of speech for each of the emotions in the EMOGIB database appear to be sufficient for many HRI studies. However when new semantic-free speech databases are generated using the framework, a shorter recording time would be desirable. The shorter the recording time the shorter the duration of the usable semantic-free speech is. Once reaching that duration limit in the implementations, if the same utterances from the database are used repetitively in the produced nonsense speech of the robotic agent, the perception of the naturalness might decrease. Thus, a degree of variation is required in the produced nonsense speech to achieve an affective communication mean for longer duration of time.

More variations of the semantic-free utterances are possible to be generated with *concatenative synthesizing techniques*, as the EMOGIB database is already in semantic-free form. Basically this would mean swapping the units of an utterance with other units from the database of the related emotion. These units are referred to as *swappable segment units* and they should share the following characteristics:

- They can be replaced by the other segments with fewer artifacts.

- They are relatively easier to be automatically detected.
- They are language independent.

Even though various phonemic units such as phrases, words, syllables, diphones, etc. could be considered as potential swappable segments, concatenation at smaller segment units are prone to more artefacts. The smaller the segments are, the more iterations of signal processing are required for the concatenation of an utterance. As each iteration can lead to some artefact, more iterations in most cases result in more artefacts. Also not all the larger potential swappable segments satisfy the characteristics lined out above. For example while concatenating two voiced segments, not only the phonemic alignment but also pitch periodicity should be assured. Also for all potential swappable segments, co-articulation and prosodic appropriateness should be considered. Based on the conformance to the above listed characteristics, the following parts of an utterance have been chosen as swappable segment units: *voiceless phoneme, part between two voiceless phonemes and part between two pauses.*

However, as prosodic and acoustic aspects, such as the pitch declination line, could be destroyed, it is not known if this kind of swapping in an utterance will damage the *emotion recognition* and *naturalness* perception. Informal experimentation has been performed for identifying potential swappable segment units and whether the modification would harm the emotion recognition and naturalness perception. The units have initially been identified as summarized in Table 5.1 and Figure 5.1 illustrates the swappable segment units with their boundaries on the same sample utterance for each of the 3 unit types.

Table 5.1: Potential swappable segment units

Unit 1	segment between two pauses
Unit 2	segment between two voiceless phonemes
Unit 3	voiceless phoneme

The preliminary results for Unit 1 and Unit 2 from these informal tests were promising for both emotion recognition and naturalness, especially once the last units of an utterance have been kept fixed and not swapped in the ordering. This created the motivation for a formal experiment, which is detailed in Section 5.2.3.

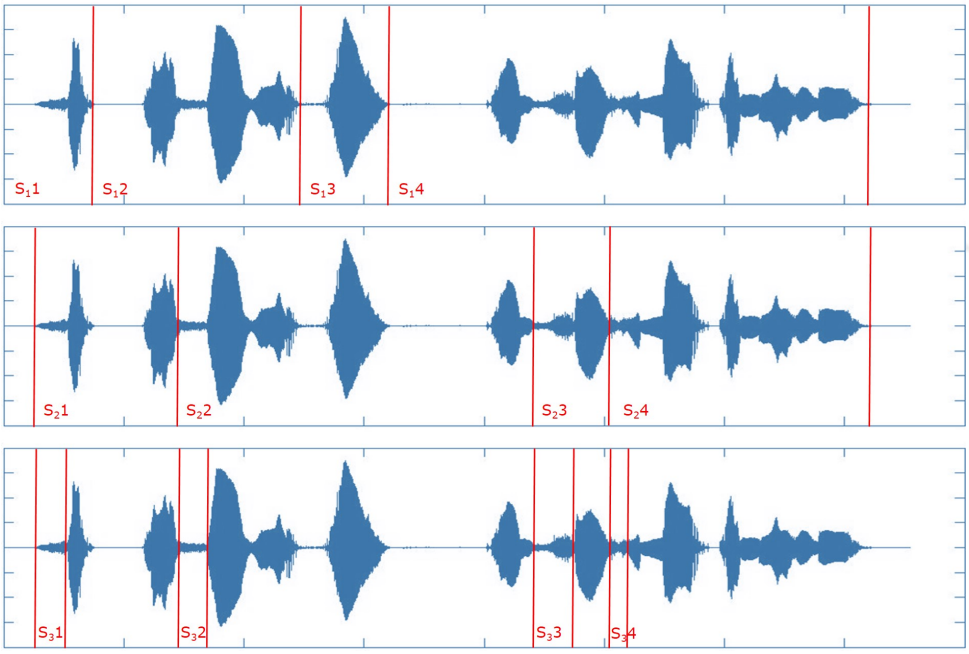


Figure 5.1: The swappable segment units with their boundaries on the same sample utterance for each of the 3 unit types. In the top figure, the swappable segments (s_1x) are the parts between two pauses (Unit 1 type), in the middle figure the swappable segments (s_2x) are the parts between two voiceless phonemes (Unit 2 type), and in the bottom figure all the swappable segments (s_3x) belong to voiceless phonemes (Unit 3 type)

Regarding Unit 3, voiceless phonemes have similar acoustic characteristics. In the informal tests, it was observed that the variation generated from swapping the voiceless segment units in an utterance was not easily recognizable. Thus giving the feeling of the synthesized utterance being the same as the original one. For this reason, Unit 3 has not been included in the final swappable segment units list for the performed experiment.

5.2.1 Database labeling

In order to be able to use a speech database for concatenative speech synthesis purposes, appropriate *meta-data* describing various aspects of the speech contained in the database needs to be generated. This includes graphemic and phonetic transcriptions, phonemic segmentation indicating phoneme boundaries, symbolic feature information such as part-of-speech, lexical stress, syllable type, etc. for different phonemic, prosodic and linguistic units, or acoustic feature information such as energy, MFCCs parameterizing the spectral information and pitch-markers indicating the pitch period in the voiced segments of the speech. Depending on the synthesizer requirements, this data is used by the synthesizer to compose the target synthetic speech. For the purpose of segment swapping by concatenation synthesis technique, having the segment boundaries (as sample points) of pauses, voiceless and voiced frames in the meta-data file is sufficient, considering there is no need for linguistic or phonemic alignment in the concatenation of gibberish speech segments.

This meta data then will be used by the synthesizer to select the appropriate original speech segments to be concatenated and compose the unique synthetic semantic-free utterances as the output.

The pause, voiceless and voiced segmentation in this study is based on energy and zero crossings. For each windowed signal frame, the *root mean square (RMS)* and *zero crossing counts* were calculated.

RMS is calculated by:

$$x_{RMS} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}|x_n|^2} \tag{5.1}$$

where x_n is the windowed speech signal frame of length N.

The *zero crossings* are determined by a sign-test of consecutive samples:

$$x_n.x_{n+1} < 0 \tag{5.2}$$

the cases matching the above condition (5.2) give the zero crossings of a windowed signal x_n and the number of zero crossings give the zero crossing count per each frame.

Voiced speech consists of mostly high amplitude damped sinusoids, where unvoiced speech consists of weak non-periodic, random-like sounds. So when the rms value is high and the number of zero crossings are low, these indicate a voiced frame. In adverse, a low rms value and a high number of zero crossings indicate an unvoiced frame. Then some selection rules were implemented to decide the length of the voiced/unvoiced segments and to improve on isolated errors. Selection rules were based on experimentally set thresholds.

5.2.2 Segment concatenation

The semantic-free speech synthesizer has to concatenate the swappable segments in a new order to construct a new unique semantic-free speech signal. As a preprocessing step the database is labeled as explained in the previous section.

The segment concatenation can be briefly described with the following steps:

Let s_i be the segment space, containing all the k swappable segments of an emotion category:

$$s_i = \{s_1, s_2, ...s_n, ...s_k\}$$

and let the template utterance U_i be a speech signal, selected from the database for the desired emotion, composed of a set of n swappable segment units from the segment space s_i:

$$U_i = s_p s_q .....s_n$$

Select a new set of n units from the corresponding emotion to form the new unique synthesized utterance:

$$U_i^{'} = s_x s_y .....s_m$$

in which, the number of segments of $U_i^{'}$ is the same as the number of segments of U_i. The segments in the selected set can come from multiple other utterances of the corresponding emotion in the database (see Figure 5.2) or they can all come from the same template utterance in a different order (see Figure 5.3).

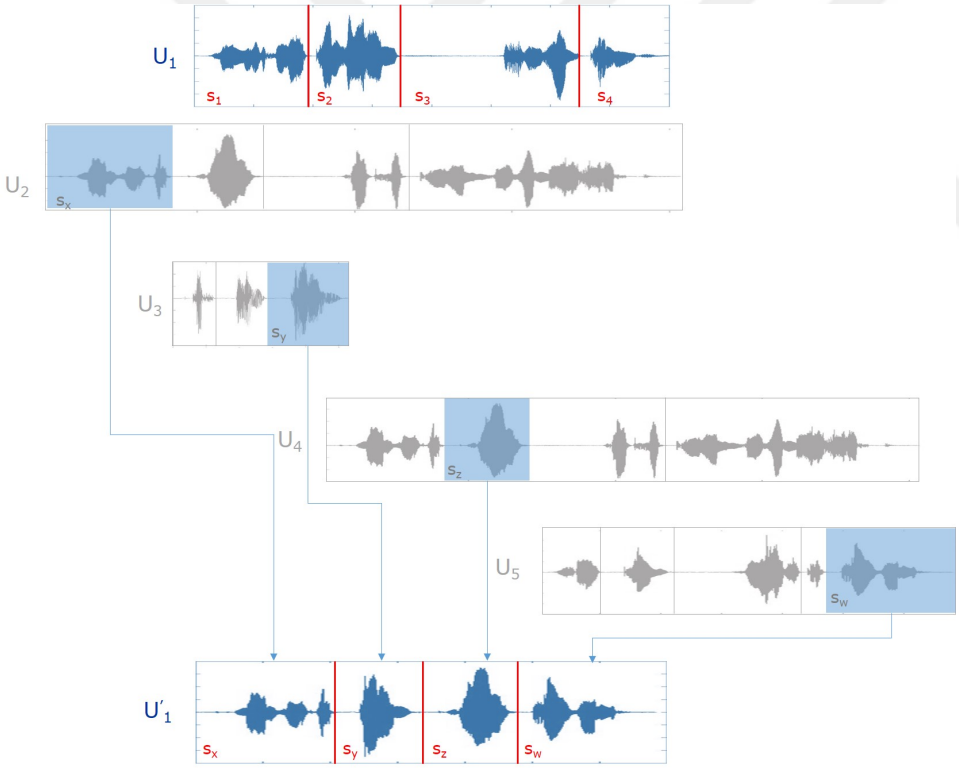


Figure 5.2: Illustration of the segment selection and concatenation for the segments coming from multiple other utterances of the same emotion database. In this example the swappable segment units are the segments between two pauses.

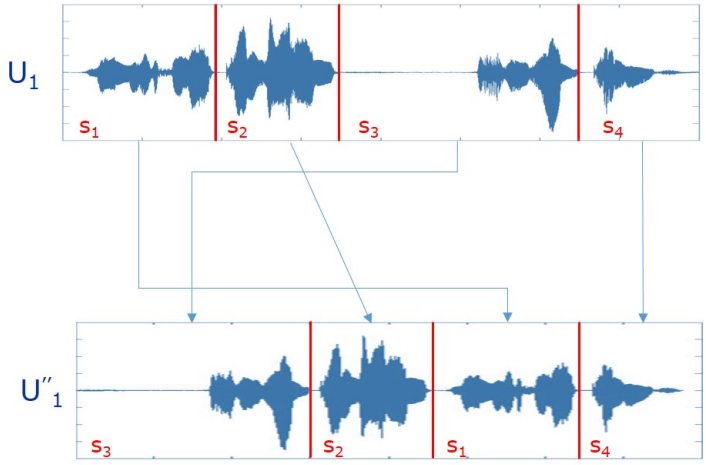


Figure 5.3: Illustration of the segment selection and concatenation for the segments (segments being the parts between two pauses) coming from the same template utterance in a different order.

This new set of n units that composes $U_i^{'}$ is selected as a random n-permutation of the k total units existing in the given desired emotion category. The number of all the possible n-permutation of the k units can be calculated as:

$$P(k,n) = \frac{k!}{(k-n)!} \tag{5.3}$$

Then each pair of consecutive segment waveforms are concatenated at segment joins. To achieve a fluent speech, the segments have to be concatenated in an appropriate way. In general this can be realized by the use of pitch markers to assure a maximum preservation of the periodicity, by pitch-synchronous overlap-add to accomplish the transition in pitch value and finally by the window/overlap operation to create the transition in waveform shapes between both segments (Mattheyses, 2013). However these steps are essential for concatenating voiced segments, especially to preserve the pitch synchronicity. In the segment swapping method described above, as one of the segment boundaries is either a voiceless or a pause, it suffices that small section of both segments that are concatenated is faded-in and out, to smooth the concatenation of two acoustic signals. Hanning window of 20ms was used for the fade-in/out operations.

5.2.3 Evaluations

The implementation potential of the two swappable segment units shortlisted in Section 5.2 (i.e. segments between two pauses and segments between two voiceless phonemes) were evaluated with an experiment. The core questions this experiment aimed to address were:

- Whether the concatenation of various segment units in a different order than their original would have an effect on *emotion recognition*
- How much the *naturalness perception* would be impacted from the concatenation of segment units and the reordering
- Which *types of segment swapping* would have the least negative effect on the emotion perception and naturalness

5.2.3.1 Stimuli

Four swapped sample sets were created: for each of the two swappable segment units; "*segments between two pauses*" and "*segments between two voiceless phonemes*", two different segment ordering schemes were utilized; "*random*" and "*fixed-end*". Summary of the sample structure can be seen in Table 5.2.

In the random ordering scheme, all the segment units of an original utterance from the EMOGIB database were reordered and concatenated randomly to create a new utterance. As a variation, the fixed end ordering scheme kept the last segment unit of an utterance in its original position while randomizing the rest of the segment units. With segment swapping, various prosodic and acoustic aspects are altered which might be negatively perceived by humans. By keeping the last segment fixed during segment swapping, some of these aspects, such as the pitch declination are less modified. By implementing the fixed end variation, it was assumed that the modifications on the prosodic and acoustic aspects would be less explicit to the listeners which would achieve a smaller negative effect on emotion recognition and perceived naturalness.

Table 5.2: Summary of the sample structure

Swappable segment unit	Ordering scheme
Segment between two pauses	random
	random with last unit being fixed
Segment between two voiceless phonemes	random
	random with last unit being fixed

For each of the two swappable segment units, a different original utterance was selected. The 2 original utterances plus the 4 swapped samples were all included in the stimuli. This sample creation procedure was repeated 4 times for each of the 6 emotions in the EMOGIB database: *anger, disgust, fear, happiness, sadness, surprise*. As such the final stimuli for this experiment included 144 samples.

5.2.3.2 Experimental procedure and participants

15 subjects (10 male and 5 female) with ages ranging between 26 and 48 participated in the listening experiment. 10 of the subjects had no prior experience with synthetic speech (naive subjects).

They were instructed to listen to various speech samples that they may not understand the meaning of and answer two questions about what they had heard for each sample.

As the first question, the participants were requested to choose which one of the given emotions matched with the speech sample they had listened to. If none of the 6 emotions matched their perception, they had the option to choose *"other"*.

In the second question, the subjects were asked to pay attention to the *naturalness* of the samples. They were instructed that a sample is considered as natural *when the sample sounds like speech of a human in an unrecognized real language, rather than sounding like an unnatural combination of vocalizations or sounds.* With this instruction it was intended to guide the users to reflect their opinion on how natural or unnatural the utterances sounded rather than judging the grammar or the content of the sentence. As such the sound of an unrecognized real language in this case is assumed to be a fluent string of speech sounds. The participants expressed their judgments using a MOS scale of 1 (Very Unnatural) to 5 (Very Natural).

The samples were provided randomly in a single presentation order. There was no time limit and the participants could replay each sample as much as they wanted. The subjects were asked to use headphones at a volume level that is high enough so the audio is clearly audible.

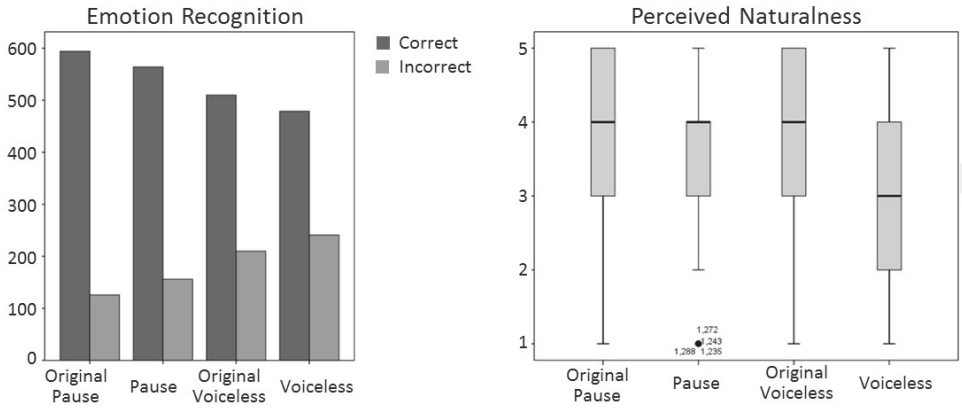


Figure 5.4: Emotion recognition rates (on the left) and the naturalness (on the right) scores for both swappable segment units and the originals.

5.2.3.3 Results

Results for swappable segment units across all emotions:

Figure 5.4, shows the emotion recognition and naturalness results for both swappable segment units (segments between two pauses – referred to as *pause*, segments between two voiceless segments – referred to as *voiceless*) and their original utterances. *Correct* stands for the emotion that was perceived was the intended emotion and *incorrect* stands for the emotion that was perceived was a different emotion than the intended one.

Regarding *emotion recognition*, analysis using Wilcoxon signed-rank tests indicated that there was no statistically significant difference between the pause segments and their originals ($Z = -1.645; p = 0.100$) or with voiceless segments and their originals ($Z = -0.927; p = 0.354$). This means that for both of the swappable segment units implemented in the framework, overall emotion recognition was not effected in a statistically significant way.

For perceived *naturalness*, an analysis using Wilcoxon signed-rank tests indicated that the naturalness scores were significantly different for each paired group. Test statistics are summarized in Table 5.3. Unlike emotion recognition, naturalness was effected in a statistically significant way for both of the swappable segment units with pause performing better than voiceless.

Drilling down to the ordering schemes, Figure 5.5 illustrates the emotion recognition

Table 5.3: Wilcoxon signed-rank test statistics for perceived naturalness

	OriPause - Pause	OriVoiceless - Voiceless	OriPause - OriVoiceless	Pause - Voiceless
Z	-7.609	-9.762	-3.695	-6.670
p	<0.001	<0.001	<0.001	<0.001

rates and the naturalness scores for the two ordering schemes (*random* and random with last unit being fixed - referred to as *fixed-end*) and their originals for segment unit *pause*.

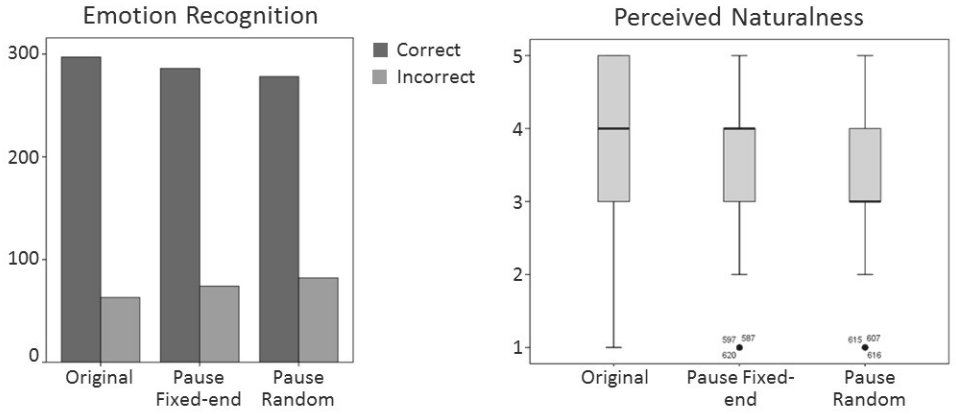


Figure 5.5: Emotion recognition rates and the naturalness MOS scores for the two ordering schemes (random and random with last unit being fixed - referred to as fixed-end) and their originals for segment unit pause

Although the Friedman test did not indicate a significant difference for *emotion recognition* ($\chi^2(2) = 4.875, p = 0.087$), it indicated a significant difference for the *naturalness* scores ($\chi^2(2) = 74.194, p < 0.001$). An analysis using Wilcoxon signed-rank tests with Bonferroni correction [1] showed that the naturalness scores of the swapped utterances were significantly different than the original for both *fixed-end* ($Z = -6.803, p < 0.001$) and *random* ($Z = -8.221, p < 0.001$) ordering schemes. On the other hand, no significant difference was found between the two ordering schemes *fixed-end* and *random* ($Z = -2.223, p = 0.026$).

For segment unit *voiceless*, the results are summarized in Figure 5.6. Again,

[1] Bonferroni correction is calculated by dividing the significance level initially being used by the number of tests that were run to gather the new significance level. In this case the new significance level is: $0.05/3 = 0.017$

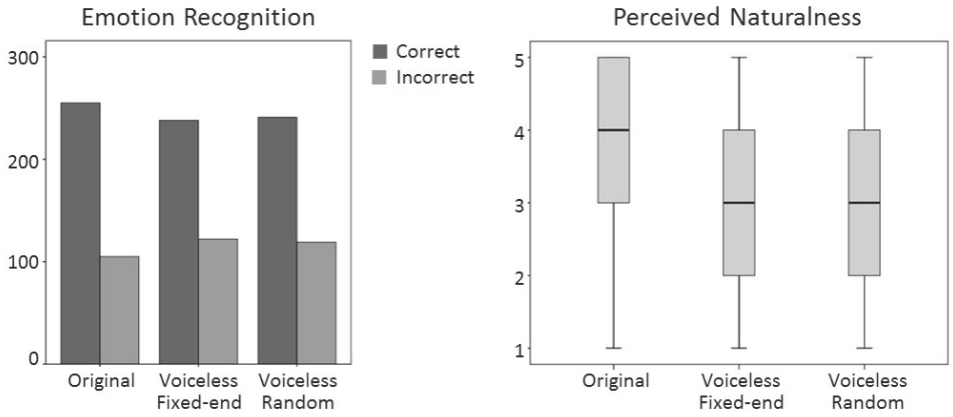


Figure 5.6: Emotion recognition rates and the naturalness MOS scores for the two ordering schemes (random and fixed-end) and their originals for segment unit voiceless

similar results were achieved as in the *pause* case. For *emotion recognition* results, the Friedman test did not indicate a significant difference ($\chi^2(2) = 3.407, p = 0.182$), while for the *naturalness* scores it did indicate a significant difference ($\chi^2(2) = 88.486, p < 0.001$). Wilcoxon signed-rank tests with Bonferroni correction showed that the naturalness scores of the swapped utterances were again significantly different from the original for both *fixed-end* ($Z = -7.695, p < 0.001$) and *random* ($Z = -9.404, p < 0.001$) ordering schemes. Also the difference between the two ordering schemes *fixed-end* and *random* was significant ($Z = -2.846, p = 0.004$) with fixed-end performing better than random.

In summary there weren't any statistically significant differences in *emotion recognition* results between pause or voiceless with their originals, as well as among any ordering schemes (fixed-end, random) implemented in each. However across each of those, the perceived *naturalness* scores were significantly different from their originals. Detailing the drill down of the results further, in next paragraph the statistical analysis per each emotion is provided.

Results for swappable segment units emotion by emotion:

Table 5.4 and Table 5.5 show the confusion matrices and Table 5.6 shows the naturalness scores for the originals of pause and voiceless samples.

Continuing on the detailed analysis, Table 5.7 shows the differences in emotion recognition and naturalness for pause or voiceless compared to their originals,

Table 5.4: Confusion matrix for emotion recognition of original pause samples (expressed in %). Rows correspond to the intended emotions and the columns correspond to the recognized emotions.

Original-Pause	ANG	DSG	FEA	HAP	SAD	SRP	OTH
ANG	**50**	7	0	3	2	12	27
DSG	0	**92**	0	0	2	0	7
FEA	5	2	**78**	0	0	10	5
HAP	3	0	3	**87**	2	3	2
SAD	2	0	2	0	**95**	0	2
SRP	0	0	0	0	0	**93**	7

Table 5.5: Confusion matrix for emotion recognition of original voiceless samples (expressed in %). Rows correspond to the intended emotions and the columns correspond to the recognized emotions.

Original-Voiceless	ANG	DSG	FEA	HAP	SAD	SRP	OTH
ANG	**58**	5	0	0	3	7	27
DSG	5	**40**	2	0	35	10	8
FEA	0	0	**62**	2	27	5	5
HAP	0	2	0	**83**	2	10	3
SAD	0	0	10	0	**87**	2	2
SRP	0	0	2	0	0	**95**	3

Table 5.6: Naturalness perception scores for original pause and original voiceless samples (on a scale of 1 - very unnatural to 5 - very natural)

Naturalness	Original-Pause	Original-Voiceless
ANG	3.8	3.6
DSG	3.7	3.8
FEA	4.0	3.9
HAP	4.0	4.0
SAD	3.9	3.6
SRP	4.2	3.9

for each ordering scheme which forms the 4 experimental groups: pause fixed-end, pause random, voiceless fixed-end, voiceless random. The table also indicates the statistical significance of each difference.

For Anger: Across four experimental groups, there were no statistically significant differences in emotion recognition compared to their originals. In perceived naturalness scores, only the drop in *voiceless fixed-end* was not statistically significant.

For Disgust: For emotion recognition only the difference in pause random was statistically significant. The difference in naturalness for *pause fixed-end* and pause random were equal and both statistically insignificant.

For Fear: Except voiceless fixed end, none of the differences from their originals in emotion recognition were statistically different. While the drops in all the naturalness scores were statistically significant, the smallest drop was in *pause fixed-end*.

For Happiness: For none of the experimental groups, the differences in emotion recognition from their originals were statistically different. For naturalness the drop in both voiceless groups were statistically significant, while for both *pause fixed-end and pause random*, there were no statistically significant differences with their originals.

For Sad: There was no statistical significance in the differences for emotion recognition, while the drop in naturalness was statistically significant in every group. The smallest drop in naturalness was observed in *pause fixed-end*.

For Surprise: Only voiceless random had a statistically significant drop in emotion recognition. All the groups had a statistically significant drop in naturalness. The least dropping group was *pause fixed-end*.

Further analysis of the test results, using Mann-Whitney U test statistics showed a significant difference between the overall scores for both the emotion recognition and naturalness of the naive subjects and speech processing experts ($Z = -3.669, p < 0.001$ and $Z = -2.952, p < 0.005$, respectively). Both the overall emotion recognition and naturalness scores of the naive subjects were significantly lower than the ratings of the speech experts.

Table 5.7: Combined differences in emotion recognition and naturalness for pause or voiceless from their originals for each emotion: difference, Z-value (p-value)

	pause_fixedend		pause_random		voiceless_fixedend		voiceless_random	
ANG								
Emo. Rec.	-10%,	-1.500 (0.134)	-3%	-0.447 (0.655)	**0%,**	**0.000 (1.000)**	0%,	0.000 (1.000)
Natur.	-0.48*,	-2.694 (0.007)	-0.77*	-3.614 (0.000)	**-0.40,**	**-2.055 (0.040)**	-0.47*,	-2.592 (0.010)
DSG								
Emo. Rec.	**-10%,**	**-2.121 (0.034)**	-19%*,	-3.317 (0.001)	12%,	-1.528 (0.127)	13%,	-1.461 (0.144)
Natur.	**-0.28,**	**-1.757 (0.079)**	-0.28,	-2.127 (0.033)	-0.72*,	-3.768 (0.000)	-0.82*,	-4.713 (0.000)
FEA								
Emo. Rec.	**9%,**	**-1.291 (0.197)**	2%,	-0.229 (0.819)	-22%*,	-3.153 (0.002)	-7%,	-1.155 (0.248)
Natur.	**-0.68*,**	**-4.158 (0.000)**	-0.77*,	-4.229 (0.000)	-0.50*,	-3.051 (0.002)	-0.73*,	3.573 (0.000)
HAP								
Emo. Rec.	**6%,**	**-1.414 (0.157)**	**3%,**	**-0.577 (0.564)**	5%	-0.775 (0.439)	2%,	-0.302 (0.763)
Natur.	**-0.17,**	**-1.287 (0.198)**	**-0.32,**	**-2.248 (0.025)**	-0.70*,	-3.536 (0.000)	-1.02*,	-4.073 (0.000)
SAD								
Emo. Rec.	**-2%,**	**-0.447 (0.655)**	2%,	-0.447 (0.655)	-14%,	-1.886 (0.059)	-2%,	-0.277 (0.782)
Natur.	**-0.45*,**	**-2.912 (0.004)**	-0.52*,	-3.098 (0.002)	-0.58*,	-3.262 (0.001)	-0.67*,	-3.990 (0.000)
SRP								
Emo. Rec.	**-11%,**	**-1.941 (0.052)**	-16%,	-2.357 (0.018)	-10%,	-2.121 (0.034)	-30%*,	-4.025 (0.000)
Natur.	**-0.62*,**	**-3.678 (0.000)**	-0.77*,	-4.522 (0.000)	-0.77*,	-3.121 (0.002)	-1.08*,	-4.167 (0.000)

5.2.4 Discussion

In general, the difference in *emotion recognition* results for both *pause* and *voiceless* swappable segment units compared to the originals were not statistically significant. The emotion recognition results of the two ordering schemes compared to their originals, which were fixed-end and random, had no statistically significant difference for both pause and voiceless swappable segment units. In other terms, segment swapping techniques implemented in the framework, didn't result in an overall drop that is statistically significant for emotion recognition. Although this was a better overall result than expected for the first core question listed in Section 5.2.3, it was not a big surprise. With the implementation of the segment swapping techniques described, many of the features, which play a role in emotion recognition, both at the frame level (e.g. raw pitch, energy) and at the utterance level (e.g. maximum, minimum, mean, range) as well as voice quality were untouched.

When the emotion recognition results were analyzed emotion by emotion, across 24 experimental groups (6 emotions x 2 swappable segment units x 2 ordering schemes) there were no statistically significant differences in 21 of them compared to their originals. *Pause fixed end* was the swappable segment unit and ordering scheme combination, which did not have any statistically significant difference across the emotions. The 3 statistically significant differences were distributed across the other swappable segment unit and ordering scheme combinations (1 per each pause random, voiceless random, voiceless fixed-end).

For *naturalness* scores, the differences for both *pause* and *voiceless* swappable

segment units compared to the originals were statistically significant. The naturalness results of the two ordering schemes compared to their originals, which were fixed-end and random, had statistically significant difference for both pause and voiceless swappable segment units. Comparing the fixed-end to random, while the difference was not statistically significant for pause, for voiceless swappable segment unit fixed-end ordering scheme performed better. Overall the naturalness was negatively affected by the segment swapping in a statistically significant way. During the segment swapping, swappable segment units are both reordered and concatenated. By the reordering performed during segment swapping, prosodic and acoustic aspects, such as the pitch contour and declination line are altered. Answering the core questions in Section 5.2.3 being addressed in this experiment also provided insights in how such an altering would be noticed by the subjects and if this would be perceived acceptable. Above mentioned results revealed that for emotion recognition the resulting alteration didn't have a negative impact while for naturalness the subjects could sense the modification hence giving lower scores. Also naturalness is prone to segmentation errors in the database labeling process. The combination of all these factors most likely have accumulated, causing the negative effect on the naturalness perception.

For the fixed-end ordering scheme, it was assumed that the modifications on the prosodic and acoustic aspects would be less explicit to the listeners which would achieve a smaller negative effect on emotion recognition and perceived naturalness. While for overall emotion recognition fixed-end and random had similar performances, in perceived naturalness in line with the above assumption fixed end performed better.

The naturalness of the segment swapping can potentially be improved by further enhancing the segment selection logic. Currently segment selection is done in a random order, except for the last segment of the utterance in the fixed-end ordering scheme. A more sophisticated segment selection logic can be introduced by defining additional rules. Such an example can be defining target and join costs which is used in most unit selection TTS synthesizers, seeking a minimum distance between the two segments to be concatenated. Considering no linguistic or phonemic alignment would be needed in the concatenation of gibberish speech segments, the algorithms in SFAS framework would most likely require fewer cost function definitions. Thus such a segment selection logic can be seen as a simplified version of the ones used in unit selection TTS synthesizers.

There were only 5 across 24 experimental groups that did not have a statistically significant difference in naturalness scores compared to their originals, when the results were analyzed emotion by emotion. *Pause* performed slightly better than

voiceless (4 vs. 1 groups with no statistically significant difference respectively) and *fixed-end* performed slightly better than random (3 vs. 2 groups with no statistically significant difference respectively).

The emotion recognition and naturalness results were also evaluated in combination. Per each emotion to identify the best performing experimental group, first the experimental groups with a statistically significant difference, compared to their originals, in emotion recognition were eliminated. For the remaining experimental groups for that emotion, if there were any which had no statistically significant difference in naturalness, these were highlighted as the best performers. If the differences in naturalness were statistically significant among all the remaining experimental groups for that emotion, the one that had the least drop was highlighted as the best performer. According to this evaluation logic, except for *anger*, *pause-fixed end* performed best across all the four experimental groups when emotion recognition and naturalness results were combined. While for *anger* voiceless fixed-end was the best performer, for *happy* both pause fixed-end and pause random were best performers. In Table 5.7 the best performing groups per each emotion are highlighted in bold.

Pause fixed-end being the overall best performer hasn't been a surprising result. Pause as a swappable segment unit, is less fragile to concatenation errors as there's no voice activity closer to the boundaries of the segment unit as long as the labeling is accurate. In combination with the fixed-end ordering scheme, as the prosodic ending was not being touched, less disturbance in naturalness was achieved.

The only exception to the pause fixed-end being the top performer in emotion by emotion analysis was anger. In emotion recognition, statistically both voiceless ordering schemes were almost one to one matching with the originals. Across all the four experimental groups, only voiceless fixed-end didn't have a statistically significant drop in perceived naturalness scores. The difference in the results for anger compared to the other emotions still to be further investigated.

5.3 Voice modification

Frequency code work of (Hinton, Nichols, & Ohala, 1994) has shown the relationship between the physical volume and the voice pitch for both humans and animals. Also the entertainment industry has been analyzing preferences of the audience for certain characters and the combination of actors' voice characteristics with the physical appearance of the characters. Such a preferred relation is also expected between robots and their voices. As mentioned in the introduction of this chapter,

there are a few notable studies focusing on the vocal attributes (Mitchell et al., 2011; Read & Belpaeme, 2010; Komatsu & Yamada, 2011). What hasn't been explored so far is finding the matching voice style for a robotic agent in alignment with the physical morphology of the robot. This gap forms up the aim of this section.

This alignment need has been explored as a side motivation during the experiment detailed in Section 6.3. This experiment was performed prior to the integration of the voice modification capabilities, which are detailed in this section, into the SFAS framework. The results from that experiment showed that the children subjects liked the voice of the robot in the experiment. However the question if the voice belonged to the robot received lower than expected scores. This have further motivated the investigations in this section on aligning the voice with the robot's morphology which can be achieved by changing the speaker's voice identity.

5.3.1 Voice modification architecture

When one speaks, two main types of information are encoded in the speech: linguistic information (message or meaning) and non-linguistic information (like speaker identity and voice quality). To change a speaker's voice identity, voice modification techniques are used so that the voice sounds like another person rather than the original speaker.

In this study, to alter the voice identity, the segmental acoustic qualities of the voice (prosodic qualities are ignored) are modified by applying a *global spectral shift* and *vocal tract modification*.

The *global spectral shift* is realized by time-scaling and resampling of the speech waveform. First, the speech signal is time-scaled using WSOLA (Verhelst & Roelands, 1993).

Ideal time-scaling algorithm is expected to produce a synthetic waveform $y(n)$ that maintains maximal local similarity to the original waveform $x(m)$ in corresponding neighborhoods of related sample indices $n = \tau(m)$ (Verhelst & Roelands, 1993). This can be expressed as:

$$\forall(m) : y(n + \tau(m)).w(n)(=)x(n + m).w(n) \tag{5.4}$$

where $w(n)$ is a windowing function, and $(=)$ stands to define 'maximally similar to'. Assuming that after Fourier transformation the maximal similarity continues,

and defining the short-time Fourier transform (STFT) $X(\omega, m)$ of $x(m)$ as:

$$X(\omega, m) = \sum_{n=-\infty}^{+\infty} x(n+m).w(n).e^{-j\omega n} \tag{5.5}$$

then the expression (5.4) can be written as:

$$Y(\omega, \tau(m))(=)X(\omega, m). \tag{5.6}$$

Once the effective length of $w(n)$ in (5.4) is selected to span at least one pitch period, the important characteristics of the signal can remain unaffected after the time-scaling operation(Verhelst & Roelands, 1993). Now, solving the time-scaling problem based on manipulation of short-time Fourier transform, gives an operational definition for (=) in expression (5.6).

Let $X(\omega, \tau^{-1}(L_k))$ represent a down-sampled version of the STFT of the input signal $x(n)$, and assume a strict equality for (=) by specifying the 2-dimensional function,

$$Y^{\beta}(\omega, L_k) = X(\omega, \tau^{-1}(L_k)). \tag{5.7}$$

The overlap-add technique in general, as well as WSOLA, proposes to synthesize a signal $y(n)$ whose STFT $Y(\omega, L_k)$ is as close as possible to the desired $Y^{\beta}(\omega, L_k)$. A tolerance is allowed on the time-warping function. WSOLA uses this timing tolerance Δ_k to ensure that the time-scale modified waveform can maintain maximal similarity to the original waveform across its segment joins in the overlap-add procedure. The basic synthesis equation, as described in (Verhelst & Roelands, 1993), is:

$$y(n) = \sum_{k} v(n - kL).x(n + \tau^{-1}(kL) - kL + \Delta_k) \tag{5.8}$$

where $v(n) = w^2(n)$ is the symmetric windowing function and kL represents ($L_k = k.L$) the synthesis instants that are chosen regularly spaced. Fig. 5.7 illustrates the operation of a basic WSOLA technique.

After time-scaling the signal with WSOLA, the modified signal is resampled to its original length. Playing back this signal at its original sampling frequency results in shifting the original signal's spectrum (Figure 5.8).

Vocal tract modifications are based on the residual-excited LPC (Linear Predictive Coding) analysis/synthesis and re-parameterization of the PLAR (pseudo log area ratio) parameter curve (Olive & Buchsbaum, 1987; Yang & Stylianou, 1998; Corveleyn, Coose, & Verhelst, 2002).

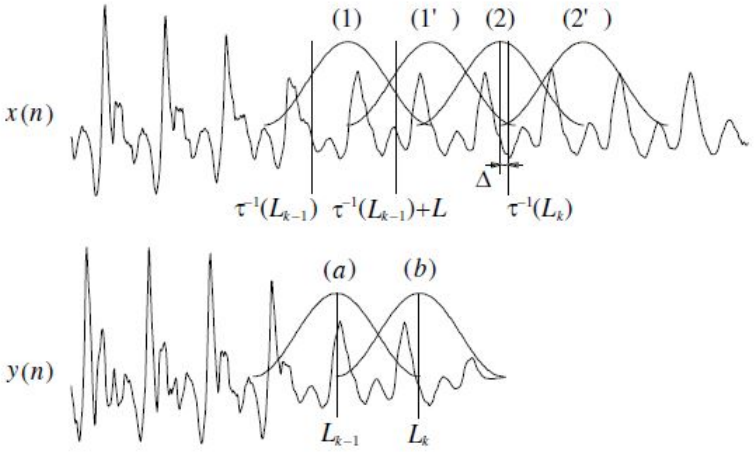


Figure 5.7: Operation of a basic WSOLA algorithm (Verhelst & Roelands, 1993). Proceeding in a left-to-right fashion, assuming segment (1) was the last segment that was excised from the input and added to the output at time instant $L_{k-1} = (k-1).L$, i.e. segment(a) = segment(1). WSOLA then needs to find a segment (b) that will overlap-add with (a) in a synchronized way and can be excised from the input around time instant $\tau^{-1}(k.L)$. As (1') would overlap-add with (1) = (a) in a natural way to form a portion of the original input speech, WSOLA can select (b) such that it resembles (1') as closely as possible and is located within the prescribed tolerance interval around $\tau^{-1}(k.L)$ in the input wave. The position of this best segment (2) is found by maximizing a similarity measure (such as the cross-correlation or the cross-AMDF(Average magnitude difference)) between the sample sequence underlying (1') and the input speech. After overlap-adding (b) with (a), WSOLA proceeds to the next output segment, where (2') now plays the same role as (1') in the previous step

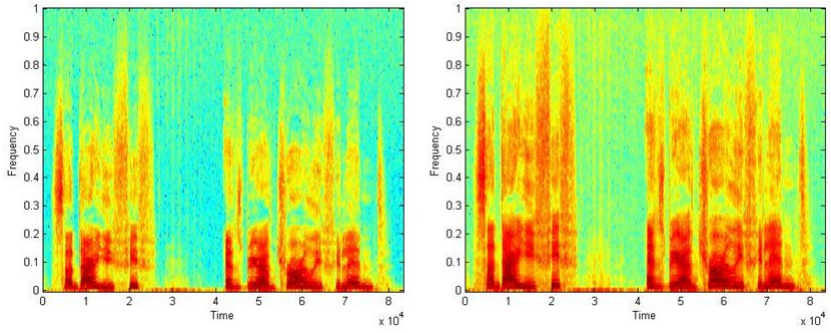


Figure 5.8: Global spectral shift is realized by time-scaling and resampling of the speech waveform. The original speech signal is on the left and the resulted signal with a spectral shift is on the right.

PLAR parameters characterize the cross-section of the vocal tract and by simply stretching or compressing some parts of the PLAR curve, it is possible to simulate the effect of a change in the length of the corresponding parts of the vocal tract (Corveleyn et al., 2002). In mathematical terms, the PLAR parameters can be defined as:

$$h_0 = 0, h_i = h_{i-1} + log(\frac{1-k_i}{1+k_i}) \tag{5.9}$$

where k_i represents the i^{th} LPC reflection coefficient. These coefficients, which are related to the transection of an acoustic tube model of the vocal tract, can be defined as:

$$k_i = \frac{A_{i+1} - A_i}{A_{i+1} + A_i} \tag{5.10}$$

with A_i the area of the i^{th} transection. Combining (5.9) and (5.10) gives:

$$h_i = log(\frac{A_1}{A_{i+1}}) \tag{5.11}$$

Plotting all the PLAR parameters and drawing a line through them by interpolation gives a visual representation (Figure 5.9). As mentioned before by simply stretching or compressing some parts of the PLAR curve, it is possible to simulate the effect of a change in the length of the corresponding parts of the vocal tract[2]. Once the warped curve is sampled at the same equidistant places along the horizontal axis to get the new PLAR parameters, the length of the warped curve may have been changed. Thus, LPC-order (the number of parameters) needs to change too, which requires a transformation operation by re-parameterization of the curve as explained in (Corveleyn et al., 2002).

Let $s(x_s)$ be the PLAR curve of the source speaker, with PLAR parameters at $x_s = 0, 1, ..., p$ (p = LPC order). Then perform the warping of the curve by applying a re-parameterization of the curve, with $x_s = m(x_t)$, which results in $s(m(x_t))$. Now, the new PLAR parameters can be found for $x_t = 0, 1, ..., p_m$, with p_m the new number of trans-sections, calculated as:

$$p_m = [m^{-1}(p)] \tag{5.12}$$

[2]If gender transformation is taken as an example, formants in female speech are higher in frequency than in male speech. This is due to the shorter vocal tract of female speakers, especially because the vocal chords are located less deeply than with male speakers. Thus, male-to-female transformation can be achieved by changing the length of the front (simply compressing such as in (Corveleyn et al., 2002)) or both front and back (compressing the front, and equally stretching the back such as in (Yang & Stylianou, 1998)) of the vocal tract.

Combining the time-scaling and resampling of the speech signal with the vocal tract conversion in the LPC domain, provides a simple and robust voice modification architecture which allows to change the initial speaker's voice identity. MATLAB user interface of the voice modification architecture described above and used in this study can be seen on Figure 5.10.

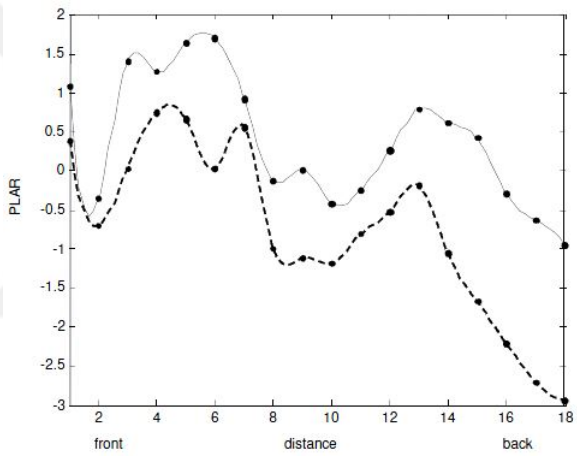


Figure 5.9: Similar sounds produced by two different speakers corresponds to two different PLAR curves (Corveleyn et al., 2002).

Additionally a real-time module that incorporates the spectral shift modification has been developed. Basically in this module the speech signal is time-scaled and the modified signal is then resampled to its original length. This module is intended for the WoZ studies where the robot is operated by and speaks through a wizard.

5.3.2 Voice alignment with the robot morphology

Various factors would have an impact on the type of voice for a given robot; such as the role of the robot (companion vs. teacher), action speed of the robot (slow vs. fast) or the social context of the interaction (partner vs. competitor). This study is focused on the impact of the physical factors (morphology) and examines the relation with the voice spectral shift.

The relation between the voice characteristics and the robot morphology was investigated in an experiment with the EMOGIB database and two robots whose

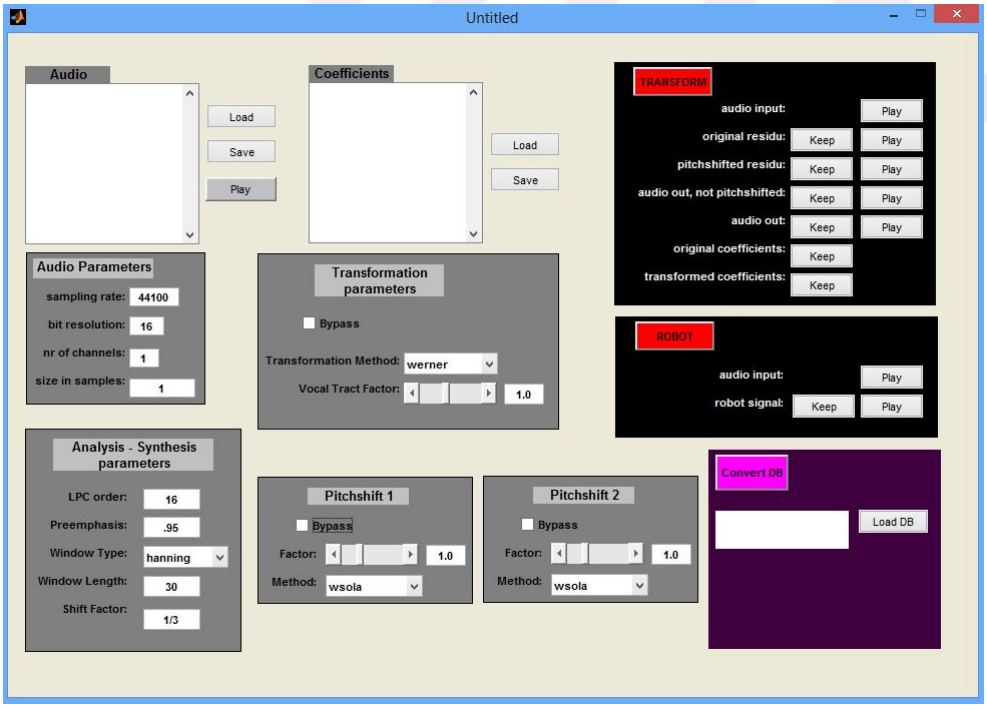


Figure 5.10: MATLAB user interface of the voice modification architecture

morphologies differ from each other. Robots Probo (Saldien et al., 2008) and Nao (*NAO 2015 - SoftBank Robotics' humaonid robotic platform*, n.d.) were utilized in this experiment. Probo is a high volume animal-like green robot with a fur and Nao is a low volume human-like gray robot with a plastic cover (Figure 5.11). These robots are the evaluation platforms used for the SFAS framework in this dissertation and more details about them will be provided in Section 6.2 of the next chapter.

	PROBO	NAO
Volume	high volume	low volume
Height	80 cm	57 cm
Type	animal-like	human-like
Color	green	gray
Cover	fur	plastic
User Group	kids	
Sex	unspecified	
Age	unspecified	

Figure 5.11: Probo on the upper-left and Nao on the upper-right with their morphology summary table below.

5.3.2.1 Stimuli

Two sets of samples were created for this experiment. For the first set, one neutral sample from EMOGIB was selected as the base utterance. From this base utterance, 4 low pitched and 4 high pitched samples were designed empirically and generated by using the voice modification technique described in Section 5.3.1. Each one of the 4 low pitched samples was created by spectrally shifting downwards compared

to the preceding sample which generated an audible gradual downward shift in the end (global spectral shift factors were: 0.784, 0.703, 0.622, 0.541). Similarly, for the high pitched samples spectral shifts were made upwards (with the factors: 1.081, 1.162, 1.243, 1.324). Fixed vocal tract conversion factors of 0.796 and 0.915 were used for the high and low pitched samples, respectively.

For the second set, a base sample composed of *neutral, sadness* and *happiness* utterances was created. This set was intended to provide the subjects an overview of each voice profile in various emotions. Again as in the first set, 4 low pitched and 4 high pitched samples were created from this base sample for producing gradual downward and upward spectrum shifts.

5.3.2.2 Experimental procedure and participants

Eight subjects participated in the experiment (with ages ranging between 28 and 33). The subjects watched a short muted video of each robot at the beginning of the test. This introduced the robots to the subjects and helped them to get familiar with their morphologies. There was a human next to the robot in each video to provide a reference about the robot's size.

The first sample set was presented in the first and second part of the test. The participants evaluated *how well* the voice samples fit Probo and Nao on a scale of 0 to 5 (0 meaning the voice doesn't suit the appearance of Probo/Nao at all and 5 meaning the voice suits the appearance very well).

In the second part of the test, one of the samples was told to be actually spoken by Probo/Nao and they were asked to guess which one was this sample.

In the third part, the subjects listened to the samples of the second set in an order from higher pitch to lower pitch. The subjects were instructed that the emotions would change during each of the samples. They were again requested to guess the sample actually spoken by Probo/Nao.

5.3.2.3 Results

As can be seen from the results of the first part (Table 5.8 and Figure 5.12), the higher pitched voice samples were perceived as a better fit for Nao (mean score of 3.4/5) while the lower pitched samples were perceived as a better fit for Probo (mean score of 3.5/5). Wilcoxon Signed Ranks Test showed that the difference between MOS of the low pitch samples and high pitch samples were statistically significant for both Nao and Probo ($Z = -5.921, p < 0.001$ and $Z = -5.919, p < 0.001$ respectively).

Table 5.8: Mean suitability scores for Probo and Nao

	Probo	Nao
Low pitch samples	3.5	1.2
High pitch samples	1.3	3.4

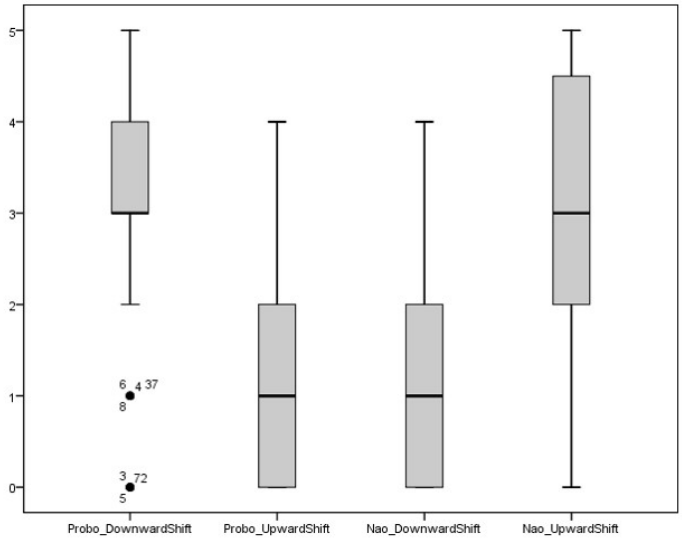


Figure 5.12: Box plots of the suitability scores of downward and upward shifted samples for Probo and Nao. The first and the second groups correspond to the suitability scores of the downward and upwards shifted samples for Probo, while the third and the fourth group show the suitability scores of the downward and upwards shifted samples for Nao.

In the second part, the voice perceived as being spoken by Nao by 63 % of the participants was the one with an upward spectral shift of 1.162. The sample with a downward spectral shift of 0.784 was selected for Probo by 50 % of the participants. All the remaining scores can be seen in Figure 5.13.

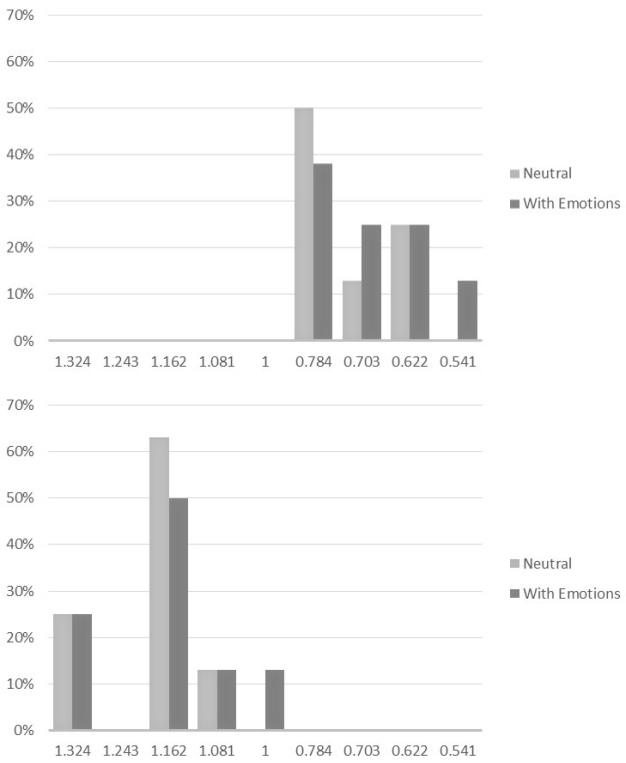


Figure 5.13: Spectral shift factor preferences for Probo (upper panel) and Nao (lower panel). Spectral shift factors are on the x-axis and the percentage of the participants on the y-axis.

The same spectral factors for both Probo and Nao were also supported by the results of the third part, but this time with a lower percentage of the participants (0.784 was selected by 38 % of the participants for Probo and 1.162 was selected by 50 % of the participants for Nao).

5.3.3 Discussion

This voice style study showed that there is a direct relation between the physical appearance of the robots and the appropriate voice pitch. As expected, similar to the findings of the Frequency Code study in humans and animals (Hinton et al., 1994), the lower pitched voices were more related with the high volume (i.e. larger) robot Probo while the higher pitched voices were more related with the low volume (i.e. smaller) robot Nao.

When different emotions were present in the voice, still the same spectral shift factors received the highest preference scores, however with a lower percentage. The fact that the emotions have also an effect on voice acoustic features, might have affected the subjects perception. For example *happiness* leads to an increase in pitch and pitch range. This creates an additional modification in the pitch on top of the spectral shift provided by the voice style modification. This might be the root cause of the lower percentage in the preference scores.

In the studies where the robot is operated by a wizard in a WoZ setup, this time the voice acoustics of the wizard will need to be aligned with the robot's morphology in real time. For example, as mentioned above if the robot being used in the WoZ set up is a low volume robot (i.e. smaller) the voice of the wizard will most likely need to be spectrally shifted upwards in real time. While this shift can be achieved using the real time implementation module described earlier, finding the most appropriate spectral shift factor will need to be experimented.

The voice alignment experiment in this study focuses only on spectral shift. Further experiments can be performed to study the relation between the physical appearance of the robots and the other parameters of the voice signal. Such a study could for instance be performed to find the most appropriate speaking rate again by using the WSOLA and re-sampling technique. Another improvement on the voice style can be achieved by timbre adjustments which can be realized by PLAR curve re-parameterization.

The motivation for choosing gibberish speech rather than other SFU types was elaborated in Section 3.1 and the results of the experiment indicate that the

subjects perceive this choice as appropriate with specific spectral shifts for both of the robots. However how the human voice based SFU would compare to other SFU types when used in a non-humanoid robotic embodiment (e.g. Probo) would be an interesting research question to explore further.

By aligning more voice parameters with the robot morphology in respect of their potential uses, it can be possible to achieve higher scores in user satisfaction in human robot interaction.

5.4 Summary

In summary, the segment swapping techniques, which are implemented in the overall framework and tested with the experiment explained above, have further enhanced the capabilities of the SFAS framework by providing a synthesizing capability with no significant negative effect on emotion recognition and acceptable levels of drop in naturalness. The major strength of this synthesizing capability is the ability to significantly increase the amount of usable and unique semantic-free utterances, without needing to perform additional recording activity. Hence segment swapping decreases the cost of implementation of the framework in HRI studies, which will most likely lead to wider and faster adoption of the framework by the HRI community.

Segment swapping significantly expands the synthesized unique semantic-free utterances. Even though each of these utterances are unique, depending on the intensity of the reordering performed, some utterances might be considered as repetitions or recurrences by the listeners. With a future study, the acceptable level of repetitions and recurrences should be explored. This will also be an important parameter in identifying the minimum recording time required to be able to build a new semantic-free speech database utilizing the SFAS framework.

For the future implementations of the framework, once the semantic-free speech database is formed, it needs to be adopted to the physical appearance of the robotic agent as suggested by the voice style study detailed in the Voice Modification section. This study showed the direct relation between the robotic agent's physical appearance and the appropriate voice pitch. The lower pitched voices are considered to be more appropriate for high volume robots while higher pitched voices are preferred for low volume robots. The voice modification implemented in the overall framework as described above will allow voice pitch to be adapted to the robot morphology in respect of their potential uses, which may lead to higher satisfaction in human robot interaction.

Some of the techniques, experiments and results mentioned in this chapter have been published in (Yilmazyildiz et al., 2012; Yilmazyildiz, Athanasopoulos, et al., 2013; Yilmazyildiz et al., 2015).

6 HRI utilizing Semantic-Free Affective Speech

6.1 Introduction

Until this point in this dissertation, the Semantic-Free Affective Speech (SFAS) framework was designed, the core hypotheses behind it were tested and confirmed, then the capabilities of the framework were further enhanced in a way that future implementations would be easier, practical and more cost effective.

In this chapter, the SFAS framework is assessed further with pilot implementations using physical robotic embodiment in multiple affective human robot interaction scenarios. Each of these pilot implementations aimed to assess different aspects of the affective HRI. In the first experiment, multi-modality and as an example the effect of using Semantic-Free Affective Speech in combination with facial expressions is assessed. The second experiment focused on a hybrid usage scenario, testing the combined use of Semantic-Free Affective Speech with Natural Language, which would further expand the implementable scenarios for the framework. During the third experiment, children subjects and the robotic agent shared the same physical space in a real life like interaction scenario, watching movie clips together. The co-viewing companion robot communicated using only Semantic-Free Affective Speech throughout this affective interaction, while the emotional context was altered to test the effects on emotion perception of children and their interaction with the robot.

These experiments aimed to test the utilization opportunity of the framework for a variety of HRI settings by piloting multi-modality, hybrid implementation and physical companion interaction cases.

6.2 The robots Nao and Probo as the evaluation platforms

Two social robots were utilized as the evaluation platforms of the proposed framework. In this section these two robots; the robot Probo (Saldien, 2009) and the commercially available robot Nao (version 4.0) (*NAO 2015 - SoftBank Robotics' humaonid robotic platform*, n.d.) are described.

The SFAS framework presented in this thesis provides a medium to study human robot interaction and is implementable to various robotic agents available. Parts of the work presented in this thesis have been utilized in and contributed to the EU FP7 funded ALIZ-E (Adaptive Strategies for Sustainable Long-Term Social Interaction) project, which focuses on the design of long-term, adaptive social interaction between robots and child users. Nao has been utilized as the research platform in ALIZ-E project. Another major project this thesis has contributed to was Project HOA16. This Vrije Universiteit Brussel funded project focused on the development of natural human/robot communication architecture and implementation of attention mechanisms, using Probo as the platform. Active participation in and contributions to these two major projects in the context of this thesis formed up an opportunity to experiment and further research the framework utilizing these two robots.

In all the experiments presented in this section, either one of the two robots were physically present in the room with subjects (Sections 6.4 and 6.5), or video recordings of one of the robots were presented to the subjects (Section 6.3).

For all of these experiments, both of the robots were programmed to behave in a manner where they exhibited natural-like behaviors allowed by their affordances. For example, Nao's LEDs in the eyes blinked and when standing the robot's weight was shifted from foot to foot; where Probo's eyelids and ears moved randomly. This was to create an *illusion of life* to prevent them being perceived as static entities.

The robot Nao

The Nao robot (Figure 6.1), is a low volume small human-like robot with a rigid plastic cover. It was designed primarily for, and currently marketed to, researchers investigating social HRI, as well as areas of science that are impacted by this (such as Cognitive/Developmental robotics). It is one of the most widely used humanoid robots for academic purposes worldwide.

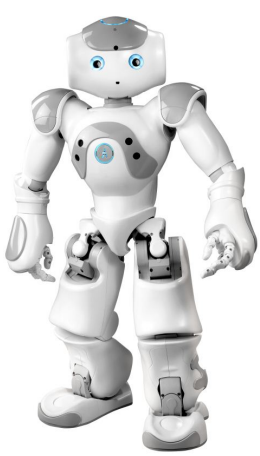


Figure 6.1: The robots Nao (on the left) and Probo (on the right) used as the evaluation platforms

The height of the robot is 57cm, the weight is 4.3kg and it has 25 degrees of freedom. The hardware properties include two loudspeakers, four microphones, two high definition cameras, a gyroscope, an accelerometer, and range sensors (2 IR and 2 sonars). The robot also has an array of Light Emitting Diodes (LEDs) that serve to represent and animate two eyes.

Nao has an embedded microcomputer (with ATOM Z530 1.6GHz CPU, 1 GB RAM) that runs a custom Linux distribution. It hosts a pseudo operating system (NaoQi) which is used to provide both a high and low level interface to the onboard resources. NaoQi also includes a built in Speech Recognition engine, Text-To-Speech engine and Computer Vision libraries (which provide onboard face detection and recognition, and object recognition).

The robot can be programmed using both the Python and C++ languages utilizing the provided Software Development Kit (SDK). As an other programming alternative, a graphical Integrated Development Environment (IDE) called Choreograph can also be used which allows programming the behaviors and utilizing on-board resources with less coding effort.

The robot Probo

The Probo robot (Figure 6.1), is a high volume animal-like robot with a green fur. It was designed as a research platform to study Cognitive Human-Robot Interaction and Robot Assisted Therapies with a special focus on children.

The height of the robot is 80cm and it has a fully actuated head, with 20 degrees of freedom, capable of showing facial expressions. It has a moving trunk and a soft huggable jacket. The rest of the hardware includes a camera, a touch screen, and a range of touch sensors.

The robot can be programmed using C# programming language in the .NET environment. A 3D model of the robot provides a real time feedback of the results. Probo is provided with a user-friendly control center which allows an operator to share control with automated systems. Also an Animation Module allows the user to assemble and manage sequences of movements/motions.

6.3 Multi-modal emotion expression

In human to human interactions, emotions in many cases are not exchanged in a unimodal way but through sets of multimodal expressions. Sharing the same social environment with humans, social robots also interact and exchange emotions with humans in a multimodal way, allowed by the affordances of the robotic agents. As such, Breazel (Breazeal, 2004) defines these various levels of affordances as one of the key differentiators between robots and synthetic agents of computer interfaces.

In essence, multimodality is a natural feature of HRI interaction with various components, such as gestural, postural, facial, auditory, etc., integrated for affective interactions. Decoding of these affects by the human users is thus dependent on the success of the combination of the affect expressions on the various levels of multimodal components.

In respect to this, a multimodal evaluation experiment was designed and performed. The effects of the Semantic-Free Affective Speech on the multimodal emotion expression of robotic agents were explored. Specifically whether the effect of the speech without semantic meaning on the emotion expression would be positive or negative when combined with another modality. In this experiment, the focus was on the two major emotion expression components of the robotic agent; auditory and facial. In a previous study, the effect of musical utterances in combination with facial expressions of an agent was researched (E.-S. Jee et al., 2007). However a similar study utilizing gibberish speech hasn't been explored.

Until this point, the emotional quality of the EMOGIB database, which was also utilized to evaluate the overall framework, was tested by adults only. Expanding the subject group variation for the evaluation of the framework, this experiment was performed with children. The social robot Probo was used as the robotic agent in this experiment. Another side motivation regarding this experiment was to test whether the children would relate the voice in the EMOGIB database to the appearance of Probo or not.

6.3.1 Stimuli

Three sets of samples were created for this experiment: visual-only sample set (V), audio-only sample set (A), audiovisual sample set (AV). Each set contained the 6 basic emotion categories, namely *anger, disgust, fear, happiness, sadness, surprise.* In the first set, video-only samples of each emotion were used without any audio. The second set consisted of audio-only samples of each emotion category without any accompanying visual modality. In the third set, audiovisual samples were used in which the audio and video samples were synchronized.

6.3.1.1 Visual stimuli

To realize a translation from emotions into facial expressions, emotions were parameterized. To construct an emotion space for Probo in this regard, two dimensions were used: valence and arousal (Figure 6.2), which was based on the circumplex model of affect defined in (Posner, Russell, & Peterson, 2005). A Cartesian coordinate system was used in the emotion space, where the x-coordinate represents the valence and the y-coordinate the arousal. Each emotion can then be represented as a vector with the origin of the coordinate system as initial point and the corresponding arousal-valence values as the terminal point. The direction of each vector defines the specific emotion whereas the magnitude defines the intensity of the emotion. Each basic emotion corresponds to a certain position of the motors to express the facial expressions (Figure 6.3) on the fully actuated head with 20 degrees of freedom (more details can be found in (Saldien, Goris, Vanderborght, Vanderfaeillie, & Lefeber, 2010; Saldien, 2009)). Using this method, smooth and natural transitions between the different emotions were achieved.

For the different visual samples of this experiment, an animation was created with the desired emotion by using the AnimationModule of Probo. To make the robot look alive when the robot wasn't performing any specific tasks, like expressing emotions, during the recordings the eyelids and ears moved randomly. The trunk and neck movements were disabled for these recordings.

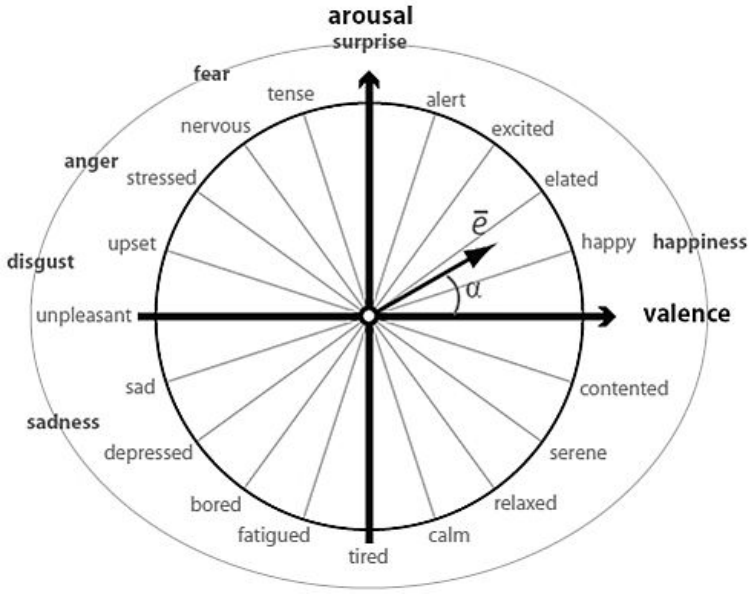


Figure 6.2: Two-dimensional emotion space of Probo, based on the circumplex model of affect defined in (Posner et al., 2005)

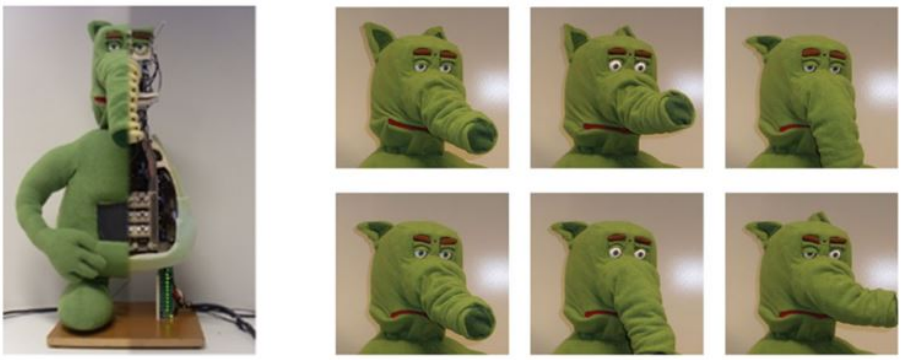


Figure 6.3: Outer and inner appearance of Probo and the 6 basic facial expressions. Top row from left to right: happiness, surprise and sadness, bottom row from left to right: anger, fear, disgust.

6.3.1.2 Audio stimuli

The audio only sample set was created by selecting random samples from EMOGIB for each emotion. The length of the samples had to be long enough so that the children could evaluate them effectively. On the other hand, the length should not be too long not to lose the attention of the participating children. Based on informal evaluations, 10 seconds was selected as the sample duration.

6.3.1.3 Audiovisual stimuli

Audiovisual stimuli were composed of audio and video samples which were synchronized. To have natural lip movements in the audiovisual samples, a lipsync module was used to match the lip movements with the speech. A commercial software package (*Annosoft Lipsync Tool 4.1*, n.d.) analysed the different phonemes from the audio file. Every phoneme then corresponded to a certain percentage of mouth opening. For example the |a| opened the mouth completely while the |m| was a phoneme that pursed the mouth the most. The two other degrees of freedom (the mouth corners) were not used for lipsync, but only for showing the emotions.

6.3.2 Participants and procedure

Thirty-five subjects participated in the test. The age range of the children was between 10 and 14. The test was performed in 2 groups of children from the same school, who were all Dutch speaking. In the first group there were fifteen children (5 male, 10 female) and in the second group there were twenty children (7 male, 13 female).

In an introductory story, all 6 basic emotions were associated with simple visual pictures. For instance the emotion *sadness* was introduced orally as a part of the story by saying and illustrating the sentence “Probo is SAD because it RAINS outside" while the picture of “an umbrella in the rain" was visually presented (Figure 6.4). During the actual testing, the children could thus associate the stimuli (i.c. Probo’s emotional expressions) with the introductory story and could tick the associated picture which is presented together with the piece of the story mentioning the name of the associated emotion in a questionnaire, even if they would not have the knowledge to semantically understand or name emotional states. This enabled the scoring of Probo’s emotional expressions (auditory, visual, and audiovisual) as perceived by the children.

The introductory story was told by a speaker who had a soft voice and a clear Dutch accent and was recorded in the recording studio (*ETRO Audio-Visual Lab*, n.d.). The speaker was instructed that he should imagine himself as a teacher telling

		
Probo is SAD because it RAINS outside. Probo wants to play outside, but has to stay inside.	*He looks through the window... then he is suddenly SCARED of a SPIDER. Probo is afraid of such big spiders!*	*Fortunately, the spider goes away quickly, and... it doesn't rain any more. Mama asks Probo to go to the seaside... Probo loves to play on the BEACH! Going to seaside is so much fun! Probo nods and is very HAPPY.*
		
Some time later Probo sits on the beach... Probo cannot go into the water... because there is a squishy JELLYFISH floating on the water? "Jellyfishes are DISGUSTING" says Probo! "Yak!" Probo cannot go into the water, so Probo builds a big and beautiful sandcastle!	*It is the most beautiful castle on the beach! Probo is so proud!* *But then the beautiful castle is just trampled by some kids! The castle is completely DESTROYED! Probo is really ANGRY and shouts that such things are really bad!*	*Because the castle is completely destroyed... mama has a SURPRISE! Probo has never seen such a great GIFT! Probo didn't expect this!*

Figure 6.4: English translation of the introductory story and the associated pictures which are also used in the questionnaire

a short story about Probo, emphasizing the emotions without acting.

All three emotional modality conditions (auditory, visual, and audiovisual) were presented to the subjects sequentially. The order of the auditory and visual modalities were counterbalanced between both groups of participants. The audiovisual condition was always presented at the end to avoid that an association between auditory and visual stimuli would affect the participants' judgments in the single modality conditions. In each modality condition, participants were asked to associate 9 stimuli in which Probo expressed an emotional state. All 6 emotional states were presented at least once. The three additional stimuli were implemented as *fillers* and were not taken into consideration in the analysis. The use of fillers aimed that participants' responses would not be affected by the strategy of excluding the already recognized emotions from the response selection. The emotion represented by the filler stimuli were randomly selected for each modality condition in each group.

At the end of the questionnaire, the participants were provided 2 additional questions. The first question requested a Mean Opinion Score (MOS) for the voice. The children had to paint a slider of 10 scales up to the point that corresponded with how much they liked the voice. The second requested a MOS score for if they thought the voice was the voice of Probo. The question was presented in the same way with a slider of 10 scales as in the first question.

The experimental sessions took around 30 minutes each. The samples were played once unless the subjects requested a repetition. This repetition occurred only a few times during the entire experiment. After playing each sample, the children were given time to fill in the questionnaire for the corresponding question. The questionnaire was prepared in a forced multiple choice structure, also including the option "I don't know".

The experiment was performed in the school's recently installed laboratory. The stimuli were presented to the subjects with a beamer on a projection screen. Prior to the actual test, the lighting conditions were adjusted for a better vision of the projection screen and also the seats were adjusted accordingly so that every child could see the screen almost equally well. Finally two loudspeakers (Alesis M1Active 520) were placed at the proper acoustic positions in the room and the sound was set to a clearly audible level for all the seat positions. The setup and one of the groups performing the experiment are illustrated on Figure 6.5.

(a)

(b)

Figure 6.5: Experimental setup (a) and one of the children groups performing the experiment (b)

6.3.3 Results

As can be seen in Table 6.1, *sadness* was always recognized best in all the modalities (60%, 100%, 97% in V, A and AV modalities, respectively). In video-only condition, it was slightly confused with *disgust*. This can partly be explained by the similarities between Probo's facial expressions of these emotions (Figure 6.3) having sligth differences on eye opening and ear movement. In the audio-only condition, fully accurate recognition was achieved which resulted in 97% accuracy in the combined modality. The confusion with *disgust* in the video-only modality was thus eliminated with the complementary information provided by the audio and a statistically significant improvement was achieved from the video-only modality to the combined modality (Wilcoxon Signed Ranks $Z = -3.357, p = 0.001$).

Table 6.1: Confusion matrix for all the modalities (expressed in %, columns represent the recognized emotions and rows represent the intended emotions)

Modality		ANG	DSG	FEA	HAP	SAD	SRP	Don't know
Visual	ANG	**34**	9	23	0	20	3	11
	DSG	11	**23**	11	9	23	0	23
	FEA	3	9	**31**	9	6	34	9
	HAP	3	0	0	**46**	23	6	23
	SAD	3	11	6	3	**60**	3	11
	SRP	6	0	31	0	0	**60**	3
Auditory	ANG	**46**	9	0	14	0	9	23
	DSG	11	**57**	0	6	9	11	6
	FEA	6	6	**71**	0	14	0	3
	HAP	34	9	3	**29**	6	6	14
	SAD	0	0	0	0	**100**	0	0
	SRP	0	3	0	11	0	**86**	0
Audiovisual	ANG	**60**	6	6	9	0	6	14
	DSG	11	**60**	0	9	11	6	3
	FEA	9	6	**69**	0	6	9	3
	HAP	3	9	6	**51**	11	17	3
	SAD	0	0	3	0	**97**	0	0
	SRP	0	3	6	3	0	**89**	0

Recognition rates for *anger* were 34%, 46% and 60% in V, A and AV modalities, respectively. In the video-only modality, *anger* was often confused with *fear* and *sadness*. This can again be explained by similar Probo facial expressions of

these emotions (Figure 6.3), especially once the trunk is excluded. In the audio-only modality, *anger* was slightly confused with *happiness*, which shares similar acoustic cues with *anger* such as increase in pitch and speech rate. While these confusions did not occur in the combined modality and the recognition accuracy was improved compared to the single modality, the change was not statistically significant with the Bonferroni correction ($Z = -2.183, p = 0.029$).

Similar improvement occurred in the *disgust* recognition results. Recognition rates were 23%, 57% and 60% in V, A and AV modalities, respectively. *Disgust* was confused mostly with *sadness*, and slightly with *anger* and *fear* in the video-only modality. When the additional information is provided with audio, in which *disgust* was slightly confused with *anger* and *surprise*, the confusion with the other emotions were decreased and the recognition rate was significantly improved ($Z = -2.837, p = 0.005$) in the combined modality. The confusion with *anger*, which existed to the same degree in both single-modalities, still remained in the combined modality.

In the video-only modality, *fear* was highly confused with *surprise* and vice-versa. This matches the findings from studies on human faces in the literature and can be partly explained because *fear* and *surprise* share similar visual cues like a wide eye-opening. In the audio-only modality, the recognition rates of *fear* and *surprise* were higher which resulted in a significantly improved recognition once combined with the video in the combined modality ($Z = -3.606, p < 0.001$ and $Z = -2.673, p = 0.008$, respectively). The recognition rates for *fear* were 31%, 71%, 69% and for *surprise* 60%, 86% and 89% in V, A and AV modalities, respectively.

In the video-only modality, the recognition rate of *happiness* was higher (46%) than in the audio-only modality (29%). In the audio-only modality, *happiness* was very often confused with *anger* and vice-versa. This can be partly explained as *happiness* and *anger* share similar acoustic characteristics of higher pitch and faster speech rate. In the combined modality the presentation of audio still improved the recognition rate (51%) but not significantly. ($Z = -0.500, p = 0.617$) Additionally, confusion with *anger* occurred very rarely while confusion with *surprise* was increased.

As can be seen from the confusion matrix, overall there was an improvement in the recognition results once the audio is provided with the video, which was statistically significant ($Z = -5.620, p < 0.001$). The AV recognition results were the best among the different modalities. Audio-only results followed closely the AV results.

Between the two groups of participants, a Mann-Whitney test did not show any significant difference ($Z = 0.657, p = 0.511$). That means the presentation order of the single modalities (A or V) did not make any difference on the combined modality (AV).

The average MOS score for if the subjects liked the voice was 7.0 (out of 10) and the average MOS score for if the subjects thought the voice is Probo's voice was 4.5 (out of 10). The Mann-Withney test did not show any significant difference either between the two groups of participants (for Q1: $Z = -1.553, p = 0.131$ and for Q2: $Z = -1.17, p = 0.254$) or between male and female participants (for Q1: $Z = -0.123, p = 0.905$ and for Q2: $Z = -0.548, p = 0.595$) for both of the questions.

6.3.4 Discussion

Multi-modal recognition test showed that the intended emotions were better recognized when the visual modality was enhanced with Semantic-Free Affective Speech. Speech without semantic information has improved affectiveness of the communication in a multi-modal setting, as it would also be expected from a natural language (Mower, Mataric, & Narayanan, 2009).

Only for *sadness* and *fear*, audio-only modality results were slightly better than combined modality results, but without a statistically significant difference (for sadness: $Z = -1.000, p = 0.3171$, for fear: $Z = -0.258, p = 0.796$).

Similar significant improvement was not seen on the audio-only condition by the presentation of video ($Z = -1.463, p = 0.144$). This might mainly be due to low recognition rates on the video-only condition. In the single modality case, the audio-only modality results were better than video-only modality results. In this study the trunk movements were excluded from the visual emotional cues, which might have contributed to the lower video-only results.

Even though audio only recognition results were promising, especially *happiness* and *anger* which were mostly confused with each other, can be further improved. This higher rate of confusion is not unexpected as they share similar acoustic characteristics like higher pitch and faster speech rate.

In Chapter 4, with the same audio corpus, much better results were achieved with adults (91% for *happiness* and 64% for *anger*). This difference can be an indication that children and adults might have a different interpretation of, especially, *happiness*.

The children thought that the voice/speech belonged to Probo with an average MOS score of 4.5, which was lower than expected. This can partly be explained by synchronization errors between the mouth openings and the speech. But it could also be influenced by the formulation of the question. The question was asked whether the children thought if it was the voice of Probo. They might have said no because they could have guessed that it was the voice of a human actor. The result could have been better if the children were asked whether they thought the voice was suited for Probo. Irregardless of the formulation of the survey question, lower than expected results for the voice suitability question formed up the motivation for the further voice alignment experiments and the voice modification module in the framework as described in Section 5.3.1. These experiments showed that the lower pitched voices were perceived as more suitable for Probo and a specific spectral shift factor on the EMOGIB voice was found desirable for Probo.

On the other hand, the average MOS score of 7.0 shows that the children liked the voice in general.

6.4 Hybrid vocal communication

In many cases interactions between people happen in combinations of multiple modalities. Considering that in their interactions with robotic agents, humans will also expect a similar multi-modality, the previous section explored the effects of combining Semantic-Free Affective Speech with facial expressions. While facial expressions and bodily gestures are most likely the first modalities coming to mind enriching the natural-like interactions in combination with speech, some forms of variations can also happen within the same modality, such as multilingualism.

Depending on the role of the robot or the social context of the interaction, in many cases Semantic-Free Affective Speech may be implemented as the sole vocal medium of the robotic agent. However, especially in cases where specific contextual information input or output during the social interaction is required, such as expressive robots as education companions or peers (Saerbeck, Schut, Bartneck, & Janse, 2010), natural language interaction will still be a vital component. When the current level of natural language processing(NLP) is considered, despite all the accelerated progress, the current implementations are still far from a state where machines are able to engage and partake in open-ended conversations (Moore, 2014; Mubin et al., 2012). As such, the implementation of Semantic-Free Affective Speech in combination with natural language can have a potential in many currently challenging scenarios in social HRI.

Another potential implementation scenario is addressing a failure in the speech recognition, TTS engine or dialog manager during an interaction scenario with this hybrid approach. In such cases, instead of producing a wrong reply statement or constraining and scripting interactions and dialogues (e.g., Lohse et al., 2008) or requesting the same input from the user multiple times (e.g., Holzapfel & Gieselmann, 2004), a semantic-free gibberish statement can help eliminating an interruption in an affective interaction. In such scenarios, Semantic-Free Affective Speech can support, instead of replacing, the natural language.

In this context, the experiment in this section intended to investigate whether semantic-free gibberish speech can be used in combination with a natural language in the interaction between a robot and humans, while assessing if such a usage is perceived as appropriate, natural and preferable or not. While a similar experiment was performed for non-linguistic utterances (Read & Belpaeme, 2014a), Gibberish speech hasn't been explored in this context before. This investigation has been performed as a part of the Social HRI Summer School in Cambridge. The social robot Nao was used as the robotic agent in this experiment.

6.4.1 Stimuli

For this experiment a game, which provides opportunities for the robot expressing various vocalisations, that is widely known as "Cups and Balls" was chosen. In the game a ball is placed under one of the three cups and then the cups are reshuffled. Then the robot guesses under which of the cups the ball is hidden. For this experiment, to ensure that the only controlled variable is the auditory output, the cups and balls game scenario was implemented in a way that the robot would always make an incorrect guess (Read & Belpaeme, 2014a).

The cups-and-balls game scenario code (which was developed in collaboration with the Centre for Robotics and Neural Systems of Plymouth University) was broken down in a number of modules, in each of which an auditory action might be performed. There were 12 modules in total (Figure 6.2). Two sets of audio stimuli were prepared for each module: natural language (English) and semantic-free speech (Gibberish).

The natural language samples were pre-recorded by using Nao's built in Text-To-Speech (TTS) engine. They were the same set of samples as in (Read & Belpaeme, 2014a). The only difference was that the voice of natural language samples were aligned with the EMOGIB database voice pitch which was selected as the best matching voice for Nao in Section 5.3.2. The inbuilt TTS voice for natural

language samples was spectrally shifted upwards to match the selected voice for Nao by using the voice modification technique described in Section 5.3.

For the gibberish samples, each module was mapped to an emotion. Then the samples were selected from the EMOGIB in accordance with the suited emotion category for each module. The natural language and gibberish samples were approximately the same in terms of duration.

The modules of the game, the scripts of the natural language, the mapping to gibberish emotion category and the durations were all summarized in Table 6.2.

Figure 6.6: Flow of the Cups and Balls game scenario

6.4.2 Participants and procedure

As mentioned before the cups-and-balls game scenario was used to investigate whether semantic-free gibberish speech can be used in combination with a natural language in the interaction between a robot and humans, while assessing if such a usage is perceived as appropriate, natural and preferable or not. To assess if the combination of SFAS with natural language would be preferred in an interaction in this exploratory experiment, the participants were requested to create auditory flows for the cups-and-balls game scenario. The resulting auditory flows could be in natural language only or gibberish only or any mixture of switching between the two.

The cups-and-balls game scenario code was broken down into modules. These modules were provided to the subjects with a graphical user interface on a com-

Table 6.2: Overview of the stimuli for Cups and Balls game

Module ID	Module name	Natural language scripts	Gibberish speech emotion category	Duration (sec.)
M01	Ball-presented	Ah, I know this game	Happy	1.6
M02	Watch mixing 1	Where's it going?	Surprised	0.8
M03	Watch mixing 2	Whoa, slow down!	Surprised	1.4
M04	Look at the user	Now then, where did it go?	Surprised	1.7
M05	Think about choise 1	Let me see	Neutral	1.5
M06	Think about choise 2	I think it's...	Neutral	1
M07	Think about choise 3	It's... It's...	Happy	0.5
M08	Point left	That one!	Happy	0.7
M09	Check Choice	Am I right?	Surprised	0.8
M10	Wrong guess	Drats!	Sad	0.5
M11	Feel bad	Oh that's a shame	Sad	1.5
M12	Crouch	I could have sworn that I was right!	Sad	1.7

puter that controls the robot. The user interface allowed the subjects to have the robot make its move, simply by pressing the corresponding button. Three options with different auditory outputs were provided for each move:

- utter natural language sample
- utter gibberish sample
- utter nothing

By going through the scenario step-by-step, each subject group had to decide in consensus which auditory action will happen after each module: continue with the same auditory output option or switch to the other one or utter nothing. The screen-shot of the user interface is illustrated in Figure 6.7.

The experiment started with an introduction in which the cups-and-balls game scenario and the graphical user interface were explained. Then each group was asked to build a vocal flow by choosing their preferred audio options for each of the 12 modules in the game scenario. They were encouraged to have a lively discussion in reaching their consensus at each choice.

The subjects sat around a table on which Nao was in the middle, with the 3 cups in front of it. The PC that is used to control Nao through the GUI was placed next to Nao. The setup can be seen in Figure 6.8. In each of the groups the subjects were divided over the two roles: operator of the robot and interaction partner. The operator of the robot used the GUI to execute the module options, while the interaction partner was playing the game with the robot. Figure 6.9 shows the experiment in action for one of the groups.

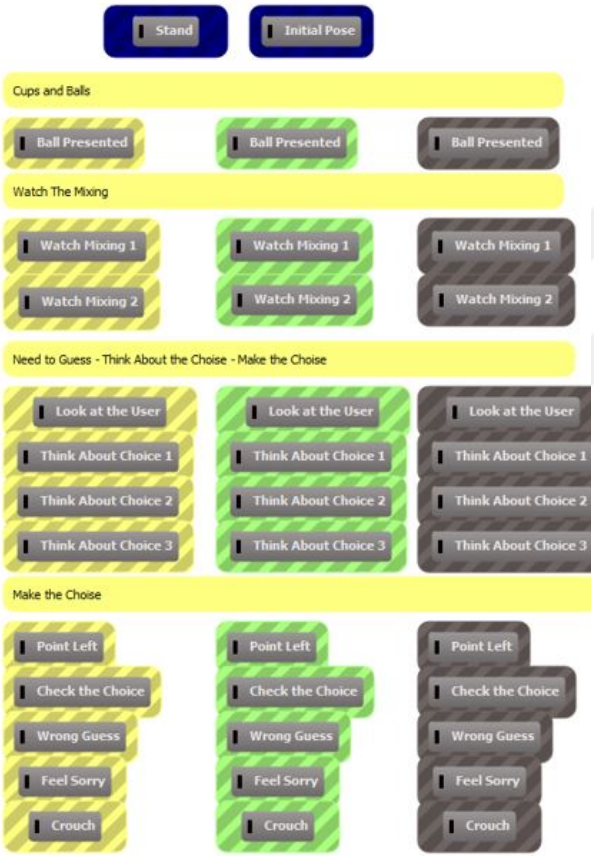


Figure 6.7: Screen-shot of the user interface used in the experiment

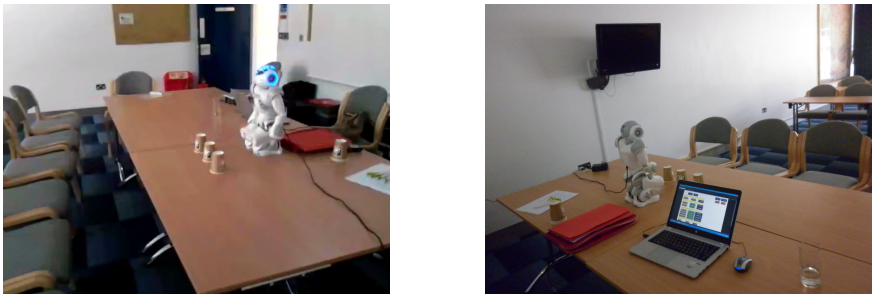


Figure 6.8: Experimental setup from frontal and backward views

Figure 6.9: One of the subject groups performing the experiment

Once the preferences were selected for the entire scenario, each group saved their flow. A video of the end to end game scenario, where Nao executed the flow of their choices, was also recorded.

The subjects were then asked to fill out a survey about their experience and particularly about the usage of gibberish in combination with a natural language. The questions queried whether the gibberish speech was able to express emotions, whether it was *appropriate* for Nao, whether it was *natural* for Nao, whether switching between gibberish and natural language (English in this case) was *appropriate* and *natural* for Nao to speak. In the second part of the survey, they were requested to give their preference ranking among 5 options: Gibberish only, English only, mixture of gibberish and English with dominantly gibberish, mixture of gibberish and English with dominantly English, and a balanced mixture. There was also an open ended question which asked for the subjects' primary strategies in choosing the auditory options in their preferred flows.

In total, there were three user groups (Group1: 9 participants, Group2: 6 participants, Group3: 10 participants). A total of 25 people attended the experiment and completed the cups-and-balls game scenario. 20 of the subjects returned their filled in survey (11 male and 9 female). It was a multicultural group with 14 unique mother tongues represented. They were all researchers in human robot interaction field and 12 of them were familiar with synthetic speech in HRI. The age range was between 23 and 35.

6.4.3 Results

While each of the auditory flows that were created by the 3 groups were unique, all the 3 groups have chosen an auditory flow that combined natural language and affective gibberish speech.

As explained before, there were 12 modules in the cups-and-balls game scenario. For three of these modules the auditory action has been chosen with a consensus of all the three groups. For eight of the modules at least two of the groups have chosen the same auditory action. Only for one of the modules, each of the three groups have chosen a different auditory action. Cronbach alpha reliability yield a moderate reliability ($\alpha = 0.677$) and Krippendorf alpha inter-rater agreement yield a weak agreement ($\alpha = 0.2064$). As the objective of the experiment wasn't to find an aligned auditory flow for the game scenario, these low alpha scores don't provide any essential statistical information for the conclusions that will be shared in the discussions in Section 6.4.4.

Table 6.3 and Figure 6.10 summarizes the preference ranking results and the average MOS scores for the expressiveness, appropriateness and naturalness of the gibberish and the hybrid usage from the survey can be seen in Table 6.4.

Although some differences in the mean values of the preference rankings can be seen in the table, none of these were statistically significant ($\chi^2(4) = 1.460, p = 0.834$). Also, there were no statistically significant differences between the male and female subjects, synthetic speech experts and non-experts, subjects who worked with Nao and those who had only seen Nao.

The subjects found the gibberish speech expressive with a mean score of 8.7 out of 10. The mean scores for appropriateness and naturalness of gibberish was 7.6 and 7.1 respectively and Wilcoxon Signed Ranks test didn't show any statistically significant difference between the two ($Z = -1.446, p = 0.148$). For switching though, there was statistically significant difference between appropriateness (mean of 7.2) and naturalness (mean of 5.8) scores ($Z = -2.149, p = 0.032$). There were no statistically significant differences between male and female subjects, as well as between subjects that worked with Nao vs. the ones who only saw it before, in the MOS scores for any of the appropriateness or naturalness for both gibberish and switching. When the statistical significance of the differences in the MOS scores of the subjects who were synthetic speech experts vs. non-experts were analyzed, the Mann-Whitney tests showed that the only statistically significant difference was in the naturalness of the Gibberish speech for Nao ($Z = -2.063, p = 0.047$), due to higher expert scores.

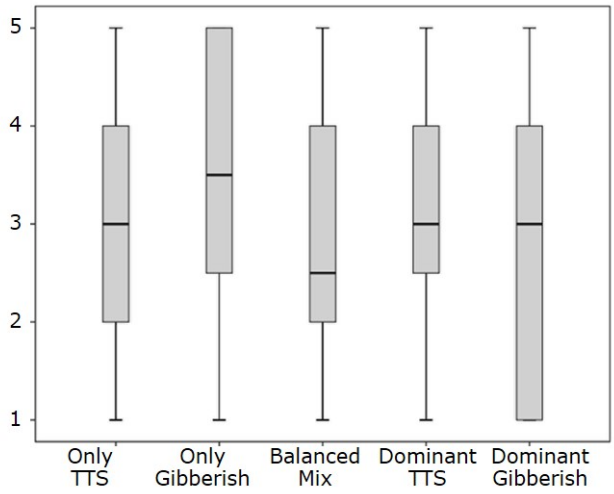


Figure 6.10: Box plots of the ranking scores for different language modalities

For any of the questions there was no significant difference between the three groups of participants.

Table 6.3: Mean ranking and resulted preference for language modalities. (1: most preferred option, 5: least preferred option)

Modality	Mean ranking	Resulted preference
only TTS English	3.2	4
only Gibberish	3.5	5
balanced mixture of TTS English & Gibberish	2.8	1
mixture of TTS English & Gibberish, dominantly TTS English	3.1	3
mixture of TTS English & Gibberish, dominantly Gibberish	3.0	2

6.4.4 Discussion

As mentioned in the Results section, each of the auditory flows that were created by the 3 groups was unique. But the result that is important for the conclusion of this experiment is that all the 3 groups have chosen an auditory flow which was switching between natural language and gibberish speech.

Table 6.4: Expressiveness, appropriateness and naturalness of gibberish and switching between natural language and gibberish

	Mean
Expressiveness of gibberish	8.7
Appropriateness of gibberish	7.6
Naturalness of gibberish	7.1
Appropriateness of switching	7.2
Naturalness of switching	5.8

Based on the feedback collected from the open ended questions in the survey, mostly the subjects preferred natural language when they felt a form of information had to be shared. When the users didn't need to understand what Nao was saying or they felt the need for expressiveness, their choice was gibberish. Such examples are Nao's response when he realizes that he made a wrong guess (M10 in Table 6.2), where all the groups have chosen gibberish or M01 when Nao provides the input that he knew the game, where Natural Language was more preferred.

In the preference rankings between only with natural language, only with gibberish, a balanced mix between natural language and gibberish, natural language dominant mix and gibberish dominant mix, no statistically significant differences were found. Also this result implies that Gibberish can be used as the sole vocal medium or in combination with a natural language in affective HRI implementations.

An interesting finding was that the MOS score for the appropriateness of the switching between Natural Language and Gibberish was much higher than the MOS score for the naturalness of the switching. In other words, the subjects indicated that even if the switching might not be considered very natural, it was appropriate for Nao to utilize it in the current interaction scenario.

As such while this experiment provides a positive indication of the appropriateness of the suggested techniques implemented using the framework, the combination of Natural Language and Gibberish in a wider variety of interaction scenarios should be tested further.

6.5 Affective interaction in a physical companion case

Gibberish speech in multimodal and hybrid settings, as explained in previous sections, aimed to contribute to social HRI, as a step toward bringing the emotional expression much closer to the way it is seen in real life. An additional step in the same line is employing Semantic-Free Affective Speech in more real-life, contextual scenarios and evaluating whether there would be perceptual changes in interpreting Gibberish speech.

This section describes a test implementation and an evaluation experiment for emotional gibberish speech with a robotic agent embodiment that is being used as one of the evaluation platforms as explained in Section 6.1. Scenarios with embodied robots in general showed more interaction (Fridin & Belokopytov, 2014), increased empathy (Seo et al., 2015), and enjoyment (Leite et al., 2008) from the users. In this scenario the robot Nao plays a co-viewing companion role in watching movie clips with children, sharing the same physical space.

Aligned with the above, the primary goal of the study described in this section was to observe how the affective interaction occurred between children and the embodied robotic agent, who would be communicating using only Semantic-Free Affective Speech, sharing the same physical space and contextual setting with the children. Secondly, measuring the changes in emotion perception of children when diverging or even confusing contextual information being provided was targeted. For these objectives, the co-viewing companion role of the robot in a movie watching scenario, provided a nice setting where the contextual information could be easily switched between various emotions, simply by the choice of the movie clips being watched.

6.5.1 Stimuli

Two sets of audio and video stimuli were prepared per emotion category: one for *in-line* and one for *confusion* cases. In the inline cases the robot expressed the same emotion with the main character in the movie clip. In the confusion cases, the robot expressed a contradictory emotion which was different than what the main character's emotion was in the movie clip. Each stimulus contained one of the 6 basic emotion categories: anger, disgust, fear, happiness, sadness and surprise.

For the audio stimuli, 2 utterances from EMOGIB were generated per emotion. The length of the samples had to be long enough so that the children could evaluate them effectively. Thus, the duration of the audio files were around 7 -10

seconds. Some additional sentences were produced in a neutral tone, to be used in greeting the child to familiarize him/her with the gibberish speech. The samples were played through the robot at predefined moments during the interaction scenario by the wizard in a wizard-of-oz (WoZ) setup.

For the video stimuli, 23 animation movie scenes were shortlisted, covering all of the emotions. Then the scenes were extracted and rated by 5 adult evaluators. The evaluators were asked to label the dominant emotion in the scenes and rate the strength of the emotion on a scale of 1 to 5. Then the scenes that had the highest ratings for the dominant emotion in the movie clips were used in the experiment. There were 12 movie clips in the final selection which are summarized in the Table 6.5. As the children attending the experiment were all Dutch speaking, the Dutch dubbed versions of the original movies were used in producing the movie clips.

Table 6.5: Final selected movie clips

Emotion	Movie	Scene
Anger	How to Train Your Dragon	Merida and Elinor fight
	How to Train Your Dragon	Merida does not want to get married
Disgust	How to Train Your Dragon	Character eats the fish that was thrown out by the dragon
	Simba	Simba eats worms
Fear	Lion King	Simba is afraid
	Lion King	Simba is chased by hyenas
Happiness	The Jungle Book	Dance
	How to Train Your Dragon	Merida happily plays with her mom
Sadness	Lion King	Simba's father dies
	Bambi	Bambi's mother dies
Surprise	Ratatouille	Gustai learns he has a son
	Up	House in the movie flies

Also four confusion cases were designed where the emotion expressed by the Semantic-Free Affective Speech of the robot was not the same as the emotion in the movie clip. A summary of the confusion cases can be found in Table 6.6.

The confusion samples were inserted amongst filler samples. The filler samples were taken from the second set of movie and audio stimuli. These were only used as fillers and not used in the evaluations.

Table 6.6: Summary of the emotions in confusion cases

Emotion label of the movie	Emotion label of the robot speech
Anger	Fear
Fear	Anger
Happy	Sad
Sad	Happy

6.5.2 Participants and procedure

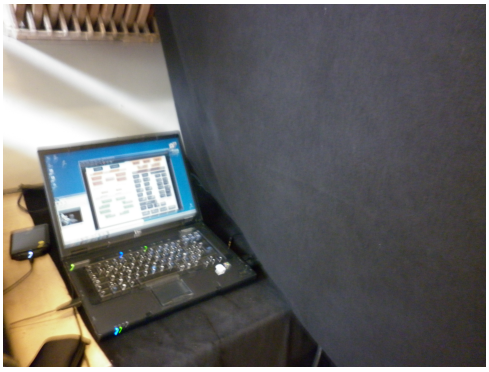
10 children attended the experiment (4 female and 6 male). The age range was between 6 and 9.

The experiment was performed at a cultural center (Beeldenstorm) as part of a "Robot Week" for children in their summer activities. All the children already had met NAO in other sessions of the robot week before they attended the movie watching experiment. To avoid any confusion with other robots, in an introductory story, the robot was presented to the children as Selo. It was told that Selo just arrived from its planet “Naoland” and thus yet only could speak its own language “Naoish”. It was learning our world and our language. The children were told that they were going to watch some short movie clips with Selo and asked to help the instructors understand what he felt about the movie. Selo sometimes could make mistakes so they should try to help him.

Considering the attention span of the children in the age range of the subjects, the experiment was split into multiple 30 minutes sessions which spanned across 5 days. As mentioned in the Stimuli section, both inline and confusion cases were represented in the samples. Two sessions per child spread across the week were designed and each session was dedicated to experiment either the inline or the confusion cases. For the confusion cases, fillers were used to ensure that the confusion samples were spread across the session. While all the children have completed the inline sessions, due to logistical challenges (e.g. child getting sick, a network outage) not all the children could complete the session focused on confusion cases. Hence not all the children rated all the confusion samples. Children could also take a break during the sessions whenever they deemed necessary.

The physical environment setup for the experiment consisted of a training area in the play zone in the garden of the Cultural Centrum and a movie chamber (see Figure 6.11) that was formed up by isolating a space in one of the rooms with black curtains.

(a) *(b)*



(c)

Figure 6.11: Experimental setup in the movie chamber: (a) two chairs were placed next to each other for the robot and the child, (b) screen was placed 2 meters away from the child and the robot, (c) wizard control of the experiment was made on a computer outside the chamber

Before coming to the movie chamber, the children one by one went through a short training session on rating along the *valence* and *arousal* dimensions using a paper and pencil version of Self Assessment Manikin (SAM) (Bradley & Lang, 1994). The training consisted of a verbal training, joint exercise with the trainer and exercise alone. In the verbal part, the trainer described the valence and arousal dimensions to the child which was followed by rating some affective pictures together. Some of these pictures can be seen in Figure 6.13. The words used to describe the poles of the valence dimension were: *happy/pleased* and *unhappy/unpleasant*. For the arousal dimension: *excited/wide-awake* were used for one end and *bored/calm* were used for the other end.

The paper and pencil version of SAM represented unlabeled dimensions pictorially on a 9-point scale. Figure 6.12 shows the SAM figures for valence and arousal dimensions. The valence scale visualized SAM smiling at one extreme and frowning at the other. A sleepy figure at the calm end of the scale and a wide-eyed excited figure at the other represent arousal.

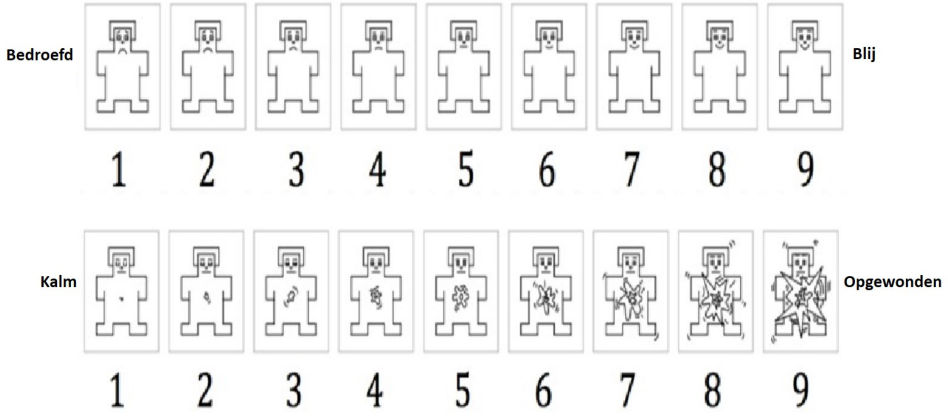


Figure 6.12: Self assessment manikins for valence (upper panel) and arousal (lower panel) used in the experiment

After the introductory story, the child was taken to the movie chamber. When the child entered the chamber, Selo greeted him/her by turning its head to the child and speaking gibberish, which was its mother tongue according to the introductory story, in neutral tone. This greeting was to make the child familiar with the robots' gibberish speech and also to make a natural start to the interaction.

Figure 6.13: Some examples of the pictures used in the training session

The child watched the movie with the robot, sitting next to each other on two chairs with an adjusted height (see Figure 6.11). The video clips were displayed on a screen. A camera was placed at a location allowing to capture frontal body streams of both the child and the robot. This allowed to record the experiment while at the same time provided a real-time streaming of the room to the wizard for monitoring during the experiment. The camera and the microphone were connected to the computer outside the room where a wizard monitored and controlled the robot as well as the overall experimental procedure. The schematic illustration of the setup can be seen in Figure 6.14. The video recordings also provided the opportunity to create a database of upper body affective expressions of the children during the interaction scenario (Wang et al., 2014).

After watching each movie clip, the robot looked at the child and uttered his emotion using Gibberish speech. Throughout the interaction scenario, when the child gazed at the robot or touched it, the robot turned his head and looked at the child. Once the interaction after watching the movie clip was over, the instructor went into the chamber with the survey and asked 4 questions to the child. First question was how *happy* the robot was after watching the movie clip. The second question asked how *excited* the robot was after the movie clip. The third and the fourth questions were asking the same dimensions but for the main character in the movie clip.

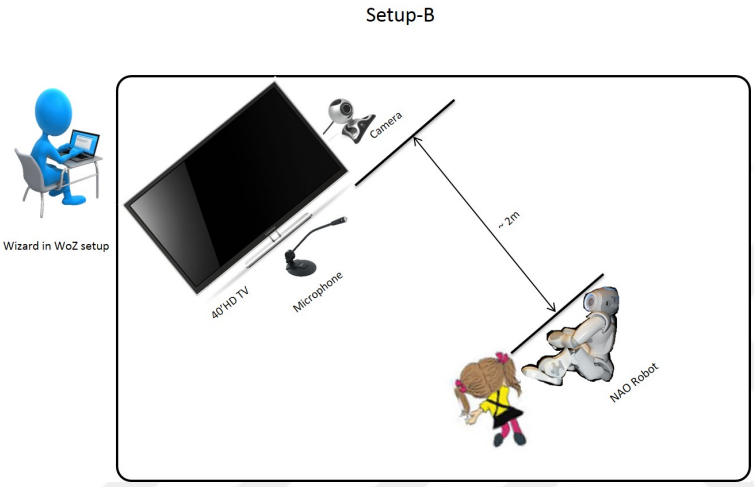


Figure 6.14: Schematic illustration of the experimental setup

After finishing the movie watching session the child had to answer 2 additional questions. The first question queried if the child wanted to watch movies again with the robot at a later time which aimed to assess the pleasantness of the interaction. The second question was to check if the child wanted to speak Naoish which could be used as an indication of the naturalness of the Semantic-Free Affective Speech and their willingness to continue to interact using it.

6.5.3 Results

Valence dimension

When the emotion in the movie and in the robot's speech were inline, it was expected that the valence ratings of the character and the robot would be correlated with each other. It was seen that indeed the valence ratings of the character and the robot were positively correlated with each other in the inline cases and this correlation was statistically significant (rs (58) = 0.375, p = 0.003).

An interesting finding on the valence degree was that the robot valence ratings were significantly higher than the character valence ratings. All the differences were significant ($p < 0.05$) according to Wilcoxon Sign test, except for happiness (see Table 6.7). The affective ratings on valence and arousal dimension for both the character and the robot are shown in Figure 6.15 .

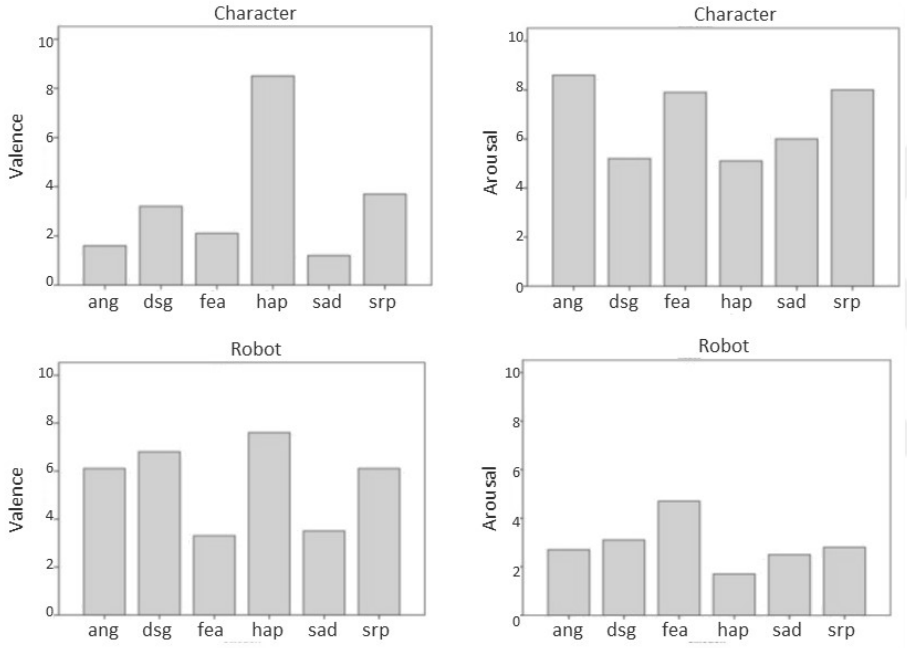


Figure 6.15: Valence and Arousal scores for the character (the two upper panels) and for the robot (the two lower panels)

Table 6.7: Test statistics of valence and arousal between character and robot in the inline case

	Valence					
	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Mann-Whitney U	-2.508	-2.677	-2.333	-0.962	-2.198	-2.205
Exact p-value	0.012	0.004	0.031	0.500	0.039	0.031
	Arousal					
	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Mann-Whitney U	-2.818	-1.703	-2.077	-2.386	-2.494	-2.680
Exact p-value	0.002	0.125	0.035	0.016	0.012	0.004

In the confusion cases, when the emotion in the robot's speech was not the same as in the movie, the valence ratings of the character in the movie and the robot were not correlated with each other (rs(14) = -0.052, p = 0.855). Furthermore, the comparison of the robot valence for each emotion between the inline case and confusion case didn't give significant difference (mean scores can be seen in Figure 6.16. and the test statistics can be seen in Table 6.8).

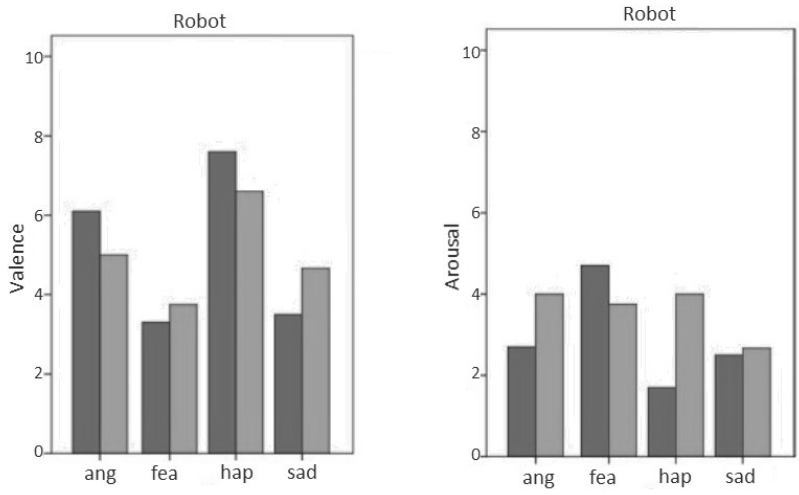


Figure 6.16: Valence and arousal scores for the inline and confusion cases

Table 6.8: Test statistics for emotion comparison

Valence				
	Anger	Fear	Happiness	Saddnes
Mann-Whitney U	10.5	17.0	20.0	12.5
Exact p-value	0.493	0.682	0.534	0.748
Arousal				
	Anger	Fear	Happiness	Saddnes
Mann-Whitney U	10.5	16.5	13.0	12.0
Exact p-value	0.476	0.665	0.098	0.699

Arousal dimension

Unlike in the valence dimension, no significant correlation was found between the arousal ratings of the character and the robot ($rs(58) = 0.079$, $p = 0.547$) for the inline cases. For the robot, the arousal ratings were lower than for the character ratings across all the emotion labels. These differences were significant for all the emotions except for disgust (see Table 6.7).

In the confusion cases, again no correlation was found between the character arousal ratings and the robot arousal ratings ($rs(14) = 0.346$, $p= 0.206$). Also no significant difference was seen between the inline cases and the confusion cases for robot arousal ratings (Table 6.8).

When the video recordings were observed, it was seen that the children showed natural interaction patterns by turning their head and listening to the robot, showing facial expressions and also talking to the robot. Some of the children even held the hand of the robot to comfort him when the case emotion was sadness. Some pictures of childrens' interaction with the robot can be seen in Figure 6.17 and 6.18.

Figure 6.17: The child pets the co-viewing companion robot's head, holds its hand and speaks to the robot during the "sad" emotion cases across multiple sessions

The question if the children wanted to learn how to speak Naoish was answered "yes" by 8 out of 10 children and the question if the children wanted to watch movies with the robot again was answered as "yes" by all the participating children.

6.5.4 Discussion

When the emotion of the character in the movie and the robot's speech are inline, it was expected that the valence and arousal ratings of the character and the robot

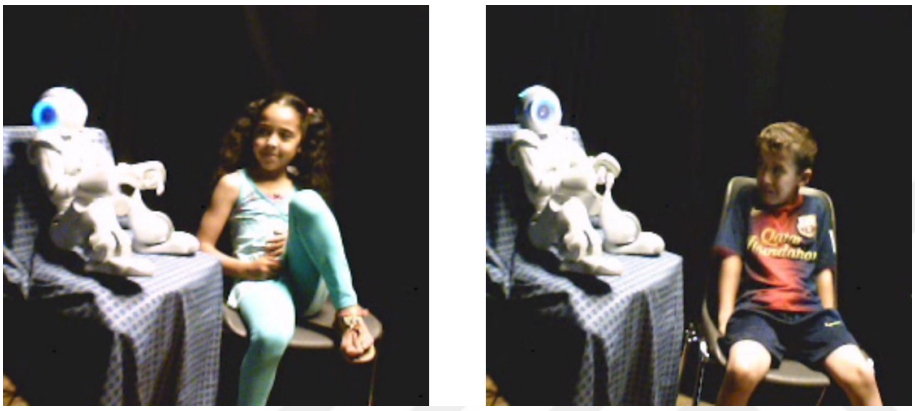


Figure 6.18: The child smiles back to the robot when the robot utters happy Gibberish speech (left panel) and the child reflects the disgust emotion in the uttered Gibberish speech with his facial expression (right panel)

would be correlated with each other. This correlation was indeed confirmed for valence in the experiments but not for arousal dimension.

An interesting finding on the valence degree was that the robot valence ratings were significantly higher than the character valence ratings (except for happiness, where no statistically significant difference was found). This might be due to the different roles the characters and the robot have in the experiment. The character is the one experiencing the affective situation in real while the robot is not in the story but just an observer watching the story. This indeed is a known effect of watching movies: enjoyment of watching negative genre movies (Neill & Ridley, 2013; Hanich, 2009; Gaut, 1993). People can enjoy watching movies having negative emotions such as fear and disgust. Moreover, the co-viewing can increase the level of enjoyment on negative genre for children. For example, in a study by Wilson and Weiss, it is found that children who viewed suspenseful movies with an older sibling liked the movie more than did those who watched alone (Wilson & Weiss, 1993). This effect might also be experienced with a co-viewing companion robot.

When the emotion in the robot's speech was not the same as in the movie, the valence ratings of the character in the movie and the robot were not correlated with each other. This suggests that the children didn't simply copy the emotion of the main character in the movie but interpreted the affective information carried by the gibberish speech, which differed from the movie's dominant emotion.

Furthermore, the comparison of the robot valence for each emotion between

the inline case and confusion case didn't give any significant difference. This also suggests that the children were able to interpret the affective information carried by the gibberish speech even when the confusing context was provided with the movie.

On the arousal dimension, the children gave lower scores for the robot than for the character. During the experiment, it was noticed that the children seemed to link the arousal with the physical "activeness" of the robot. As only one degree of freedom (head right-left) was used in the experiment and physical activeness was low, the children might have given a low rating for the arousal dimension for any of the emotion labels. The differences in the arousal ratings for the robot compared to the main character of the movie clip were significant for all the emotions except for disgust.

Unlike in the valence dimension, no correlation was found between the arousal ratings of the character and the robot for the inline cases. It can be argued that the lack of physical "activeness" of the robot might also have an effect in the missing correlation for the inline cases. In the confusion case, like in the valance dimension, no correlation was found between the character arousal ratings and the robot arousal ratings. Also, no significant difference was seen between the inline case and the confusion case for robot arousal ratings.

In summary, it is seen that the children gave higher scores on the valence dimension and lower scores on the arousal dimension for the robot than for the character. When the emotion in the movie and in the robot's speech are inline, the valence ratings of the character and the robot were correlated with each other but no such a correlation was found in the confusion case. Thus the results suggest that, the children didn't simply followed the character's emotion but were able to distinguish that there was a difference between the movie characters emotion and the emotion expressed by the robot in valence dimension. Most likely due to the lack of physical "activeness" of the robot, a similar correlation didn't exist in arousal dimension for the inline cases.

8 out of the 10 children attending the experiment said that they would want to learn how to speak "Naoish", the language the robot was speaking. This can be viewed as an indication that the children have found the Semantic-Free Speech, as the auditory output of the robot, natural in this co-viewing companion case. Also the fact that they all were willing to watch movies with Nao again in the future indicates that the co-viewing interaction with the companion robot, who was speaking Semantic-Free Speech only, was pleasant for the children.

Overall the results from this co-viewing interaction scenario with Semantic-Free Speech, which was implemented using the SFAS framework, were encouraging for future implementations. This was the first experiment where the children and the robotic agent physically shared the same space and interacted through an audio channel implemented using the framework. Despite the lack of physical "activeness" of the robot aside from his head movement in the overall interaction scenario, the children showed natural interaction patterns by turning their head and listening to the robot, showing facial expressions and also talking to the robot. Some of the children even held the hand of the robot to comfort him when the emotion of the case was sadness. Also throughout the interaction scenario, when the child gazed at the robot or touched it, the robot turned his head and looked at the child. In these cases when the robot also responded by uttering a short Gibberish expression in the same emotion as the movie clip, it was observed that the children were even more expressive.

The children were able to distinguish the differences in emotions that were transferred in semantic-free gibberish by the co-viewing companion robot, even when the contextual setting was confusing or complicated. They felt the affective interaction with a robot that was speaking only Gibberish in this scenario, was natural and pleasant. While the communication of the emotions was found to be quite effective in the valence dimension, the lack of correlation between the robot and the character for the inline cases in the arousal dimension makes it difficult to interpret the potential effect of various contextual parameters (e.g. the effect of co-viewing in emotion interpretation, the lack of physical "activeness" of the robot, etc..). It is also uncertain if the children related robot's perceived emotion with the emotional state they were in themselves after watching the movie clip. Collecting children's emotions as a data point would have helped in providing more insights, especially for the arousal dimension and validating the potential effect of the lack of physical "activeness" of the robot.

6.6 Summary

In this chapter, piloting the implementation of the outlined Semantic-Free Affective Speech Framework, sets of experiments that assess effectiveness of using semantic-free gibberish speech across various aspects of affective human-robot interaction were performed.

The pilot implementations of the framework with gibberish speech in multi-modal setting and in hybrid mode along side a natural spoken language, aimed to bringing the emotional expression closer to the way it is seen in real life. Also

Semantic-Free Affective Speech was employed in a real-life like, contextual scenario interacting with the children to evaluate whether there would be perceptual changes in interpreting semantic-free Gibberish speech.

In all the experiments presented in this chapter, either one of the two robots used as the evaluation platforms were physically present in the room with subjects or video recordings of one of the robots were presented to the subjects.

In the first experiment, multi-modal recognition tests performed with children showed that the intended emotions were better recognized when facial expressions were enhanced with Semantic-Free Affective Speech. Speech without semantic information has improved affectiveness of the communication in a multi-modal setting, as it would be expected also from a natural language.

The second experiment's results implied that Semantic-Free Affective Speech can be used as the sole vocal medium or in combination with a Natural Language in affective HRI implementations. This conclusion further expands the applicability of the SFAS framework in various real-life scenarios where the robotic agents might be playing different roles. The subjects also indicated that even if the switching might not be considered very natural, it is considered appropriate when utilized by the robotic agent in a specific interaction scenario.

One of the goals of the study described in the third experiment was to observe how the affective interaction occurred between children and the embodied robotic agent, who communicated using only Semantic-Free Affective Speech, sharing the same physical space and contextual setting with the children. Another goal was measuring the changes in emotion perception of children when diverging or even confusing contextual information was provided. The children were able to interpret the differences in emotions, that were transferred with Semantic-Free Affective Speech by the co-viewing companion robot, even when the contextual setting was confusing or complicated, but mainly in valence dimension. The potential effect of the lack of physical "activeness" of the robot in this pilot implementation to the arousal dimension requires further experimental validation. The children showed natural interaction patterns with the robot throughout the interaction by touching, smiling and speaking to the robot. The children felt the affective interaction with a robot that was speaking only Gibberish in this experiment, was natural and pleasant.

The results of these three experiments show the expansive applicability of the proposed Semantic-Free Affective Speech Framework in social HRI.

Some of the techniques, experiments and results mentioned in this chapter have

been published in (Yilmazyildiz, Henderickx, et al., 2013; Yilmazyildiz et al., 2015; Wang et al., 2014).

7 Conclusions

7.1 Summary and conclusions

Recent developments in robotics, artificial intelligence, and machine learning is further accelerating the introduction of robots in our daily lives and the physical environment around us. Increased sharing of the same physical space with humans in real life has important implications in the design of the robots. These implications are not only in the physical or mechanical design but also in the behavioral patterns of the robots. In order for humans and robots to become cohabitants, robots should behave and operate in ways that are similar to or acceptable by humans. This means they also need to be social, as humans are.

The field of social Human-Robot Interaction (sHRI) focused on the design, development and study of these socially capable agents, empowering them with a variety of social cues that allow them to interact and communicate with people in natural and intuitive ways. These not only include the use of bodily and facial gestures, natural language, and eye gaze but also more unique and robot specific methods such as expression through colors, synthetic sounds and vocalizations.

Even though the Natural Language Interfaces (NLI) have been one of the ultimate ambitions of Human-Machine Interfaces since a long time and this area has been heavily researched, the current state of the art in these technologies as a whole is still far from a state where machines are able to actively engage and participate in open-ended conversations. In parallel the demand for development and deployment of social robotic systems that should interact with people is increasing so rapidly that the rate of advances in NLI are not able to catch up yet. Not to hold back the acceleration of development and deployment of social robotic agents in real life scenarios, alternative strategies that could complement or in certain implementation scenarios that could replace NLI might be required.

In addressing this need, a framework that allows to study affective human-robot interactions by using vocalizations that do not involve semantics of a natural spoken language has been detailed in this thesis. First, the strategy of creating this Semantic-Free Affective Speech (SFAS) Framework was described along with the

evaluations on isolated audio utterances. That was then followed by pilot implementations of the framework to real world social robots and providing insights on the potential usage while seeking answers to questions related to its further deployment.

Semantic-Free Affective Speech, which is also referred to as Gibberish speech, is a member of the umbrella concept of Semantic-Free Utterances (SFU) that is introduced in Chapter 2. SFUs are human-like vocalizations and computer-generated nonvocal sounds and sound effects. Despite the commonalities in their objectives regarding social HRI, until now these auditory interaction modalities other than natural language were not investigated under a single umbrella. With this comparative study, the need for a comprehensive study of the existing literature for SFUs is addressed, the current grand challenges and open questions are outlined. Also multiple promising but currently understudied areas of SFUs have been identified as a guideline for future researchers (e.g. contextual setting of the HRI, multi-modality).

Contributing to this young field in sHRI, the Semantic-Free Affective Speech Framework, which allows robots to express and communicate through vocalizations of meaningless strings of speech, *provides a complete set of tools that can be used as a vocal communication medium for a robotic agent, which also allows to study diverse aspects of affective human-robot interaction.*

As the first component of the SFAS framework, a semantic destruction strategy was developed in Chapter 3. The semantics of an existing text in a natural language was destroyed by replacing the vowel nuclei and consonant clusters of the text using a weighted swapping mechanism in accordance with their natural probability distribution in the same language. Also the affective charging capabilities and the naturalness of the resulting semantic-free text was tested in multiple experiments utilizing TTS engines. The experimental evaluations showed that *gibberish speech resembled a natural language and it communicated the intended emotions as effectively as semantically neutral speech.*

However, in these tests, subjects reported that the synthesis engine quality affected their evaluations. In this regard, the final expressive speech strongly depends on the TTS engine quality. The quality of the expressivity models of TTS engines at the time were not yet mature enough. Also the voice quality of the emotions was not fully transmitted to the synthesized speech using the currently available TTS engines. To overcome these drawbacks in the SFAS framework to a large extent, a data-driven method was developed. As a first step of this data-driven method in Chapter 4, an emotional gibberish speech database (EMOGIB) was built and also was made available to the HRI community for further research.

It was seen that *the Semantic-Free Speech created with the data driven method resembles a natural language.* Also the perception experiments showed higher emotion recognition results for the EMOGIB database in comparison to earlier results in the field. Combining these positive results from the adult and children experiments, it was also concluded that *the EMOGIB database, as a core component of the SFAS framework, can be used in further studies across subjects with various age groups.*

Once the fundamentals of the SFAS framework were designed and the core hypothesis behind it were tested and confirmed, two modification techniques that are instrumental to the further utility of the SFAS framework in social HRI were explored in Chapter 5: segment swapping and voice modification.

Segment swapping focused on the concatenative synthesizing mechanisms to expand the number of unique semantic-free utterances in the EMOGIB database. This technique was implemented in the SFAS framework, without a significant negative effect on emotion recognition and acceptable levels of drop in naturalness. The major strength of this capability is the ability to significantly increase the amount of usable and unique semantic-free utterances, without needing to perform additional recording activity. Hence *segment swapping decreases the cost of implementation of the framework in HRI studies which will hopefully lead to wider and faster adoption of the framework by the HRI community.* The results of the segment swapping study in the future can also provide a guidance on the optimal recorded speech duration for researchers who would prefer to generate their own expressive Semantic-Free Affective Speech databases utilizing the SFAS framework.

These expressive Semantic-Free Affective Speech databases are mostly intended to be utilized through robotic agents. Robotic agents also have various physical attributes that complement their physical morphologies. The effects of misalignments between these attributes and the robot's physical morphology is fairly well studied in the field. What hasn't been explored in the literature so far and which is being addressed in the voice modification study is, finding the matching voice style (specifically voice spectral shift) for a robotic agent in alignment with its physical morphology.

This study showed the direct relation between the robotic agent's physical appearance and the appropriate voice pitch. The lower pitched voices are considered to be more appropriate for high volume robots while higher pitched voices are preferred for low volume robots. *The voice modification implemented into the SFAS framework, allows the voice pitch to be adapted to the robot's morphology, also in*

real-time. This feature may potentially increase the satisfaction in social human robot interaction for future implementations of the SFAS framework.

Once the capabilities of the SFAS framework were further enhanced in a way that future implementations of the framework would be easier, practical and more cost effective, the framework was then assessed further with pilot implementations using physical robotic embodiment in multiple affective human robot interaction scenarios. Each of the three pilot implementations detailed in Chapter 6 aimed to assess different aspects.

With various components, such as gestural, postural, facial, auditory, etc. integrated for affective interactions, multimodality is a natural feature of sHRI interaction. Human interaction partners' ability to decode emotions is dependent on the success of combining affect expressions across utilized multimodal components. In the experiment performed in Section 6.3, the effect of the speech without semantic meaning on the emotion expression combined with another modality, which was facial expressions as a visual modality in this case, was assessed. The multi-modal recognition tests performed with children showed that the intended emotions were better recognized when the visual modality was enhanced with Semantic-Free Speech. *Speech without semantic information has improved affectiveness of the communication in a multi-modal setting*, like it would also be expected from a natural language.

Depending on the role of the robot and the social context of the interaction, in many cases Semantic-Free Affective Speech may be implemented as the sole vocal medium of a robotic agent. However, in cases where the social interaction scenario requires specific contextual information input or output, natural language could still be a vital component. When the current level of natural language processing (NLP) is considered, despite all the accelerated progress, the current implementations do not fully satisfy the needs of sHRI. As such, the implementation of Semantic-Free Affective Speech in combination with Natural language can have a potential in many currently challenging scenarios in social HRI. In this context the hybrid vocal communication study in Section 6.4 implied that *Affective Semantic-Free Speech can be used as the sole vocal medium or in combination with a Natural Language in affective HRI implementations*. This hybrid usage potential further expands the applicability of the SFAS framework to various real-life scenarios where the robotic agents might be playing different roles.

SFAS in multimodal and hybrid settings, as summarized above, aimed to contribute to social HRI, as a step toward bringing the emotional expression much closer to the way it is seen in real life. The final experiment in Section 6.5 focused

on employing Semantic-Free Affective Speech in a more real-life and contextual scenario. The primary goal of this experiment was to observe how the affective interaction occurred between children and the embodied robotic agent, who communicated using only Semantic-Free Affective Speech. The robot shared the same contextual setting and physical space with the children. Secondly, when diverging or even confusing contextual information was provided, the changes in emotion perception of children were measured. For these objectives, co-viewing companion role of the robot in a movie watching scenario was used. This scenario provided a nice setting where the contextual information could be easily controlled and switched between various emotions.

Overall results from this co-viewing interaction scenario were encouraging for future implementations of the SFAS framework. In the valence dimension, the children were able to distinguish the differences in emotions, that were transferred with Semantic-Free Affective Speech by the co-viewing companion robot, even when the contextual setting was confusing or complicated. According to the children subjects, *the affective interaction with a robot that was speaking only SFAS implied to be natural and pleasant.*

The results of these three pilot implementations indicated the expansive applicability potential of the proposed SFAS framework in social HRI.

7.2 Perspectives for future work

The evaluations of the SFAS framework both on isolated audio utterances and in pilot implementations utilizing the physical robotic embodiment have shown that the quality of the overall framework is already at an implementable level. However, considering that the existence of the SFU research in sHRI is a young field, there are many opportunity areas and aspects to explore to further improve various components of the framework.

This section outlines a number of potential future directions that are considered valuable for further improving the Semantic-Free Affective Speech Framework. These potential future work are grouped under 3 main categories: language and cultural aspects, quality enhancements, long-term HRI.

7.2.1 Language and cultural aspects

While there are common tendencies in the way emotions are interchanged across different cultures (Abelin & Allwood, 2000), it's not certain how strong the role of the mother tongues and cultures of the people are, especially in social human robot

interactions.

As outlined in Chapter 2, in the current literature, the majority of the studies concerning SFUs were performed within cultural settings. Although some of the evaluations were performed with participants coming from multi-cultural and multi-language backgrounds, no real cross-cultural analysis have taken place. Also in this dissertation, most of the experiments performed included subjects from multiple mother tongues and cultures. But no deep cross-cultural experimentation and analysis utilizing the SFAS framework was done.

Assessing the acceptability and performance of Semantic-Free Speech in affective human robot interaction scenarios across various cultures could be an important next step. In the unlikely case of the results of this assessment showing significant culture/language dependence, a number of potential enhancements to adopt the SFAS framework to this revised requirement are foreseen.

Current implementation of creating Semantic-Free text in Chapter 3 takes into account only consonant/vowel distribution. However there is a lack of one to one mapping between phonemes and graphemes, which might be important for some languages. For example in English the end of the words "sandwich" and "language" sound the same but they are spelled completely differently.

Considering the above, the perceived naturalness of the resulting gibberish text and the recognition of the initial languages could be further improved by using a swapping mechanism in the phonetic transcriptions, which is closer to the speech than the text. Such a swapping mechanism would first convert the graphemes into phonemes, then would transform the result into semantic-free utterances using phoneme frequencies and then would convert those back to text to be utilized in recording enhanced versions for EMOGIB, or could synthesize speech directly using the phonetic input.

7.2.2 Quality enhancements

The shortcomings of the currently available TTS engines at the time of this study were effectively overcome by the utilization of the data-driven method in SFAS framework. Considering the attention Natural Language Processing is currently getting with the recent developments in artificial intelligence, machine learning and deep learning, future TTS engines may potentially satisfy the requirements for synthesizing the semantic-free affective speech. To validate this potential, a scanning of the state of the art affective TTS engines and evaluation of these versus the requirements of SFAS framework should be evaluated. This validation and

evaluation can be performed periodically.

At short term, the likelihood of identification of an affective TTS engine that would satisfy all the requirements might still be low. In that case, the synthesizing quality of the SFAS framework can still be further improved by enhancing the naturalness of the segment swapping. One of the potential enhancements in this regard is to improve the segment selection logic. Currently segment selection is done in a random order, except for the last segment of the utterance in the fixed-end ordering scheme. Additional rules in segment selection process can be introduced. Such an example can be defining target and join costs in segment unit selection, which seeks a minimum distance between the two segments to be concatenated. This type of segment selection logic can be seen as a simplified version of the segment selection algorithms utilized in most unit selection TTS synthesizers. Segment selection algorithms in SFAS framework would most likely require fewer cost function definitions as there is no need for linguistic or phonemic alignment in the concatenation of gibberish speech segments.

Also the precision of the labeling of segment boundaries plays an important role in the perceived naturalness of the concatenated segments. As such another improvement opportunity that may have a positive effect on the naturalness of segment swapping is increasing the accuracy of the database labeling.

The main objective of the segment swapping implementation is to expand the number of synthesized unique semantic-free utterances. While each of these synthesized utterances are different from each other, some utterances might be considered as repetitions or recurrences by the listeners in case of two utterances having a large number of common segments in a similar order. Eliminating such repetitions and recurrences in an algorithmic manner is not a complex task, once the acceptable level of repetitions and recurrences is a known parameter. Thus the acceptable level of repetitions and recurrences by the users should be explored in a future study. This will also be an important parameter in identifying the minimum recording time required to be able to build a new Semantic-Free Affective Speech database utilizing the framework.

While talking in a certain emotion, people don't only use speech to express emotions but also produce various nonspeech vocal sounds for different emotions. For example "Yuk!" for disgust, "hiii?" for surprise, or "rrrrrr" for anger, etc. These sounds, also referred as 'interjections', are short sounds heavily charged affectively that humans use naturally when expressing certain emotions. Thus, inclusion of these sounds into the EMOGIB database, or even identifying some of the segments in EMOGIB which match these characteristics and utilizing them accordingly as

interjections, may potentially increase the naturalness of the speech and the perception of the emotions from the Semantic-Free Affective Speech. Some exploratory work was already conducted to investigate this potential improvement.

7.2.3 Long-term social human robot interaction

Robots are making their way into the daily lives of the wider population in an accelerated speed. As the robots actively engage and take part in more and more real life scenarios, they interact with humans more frequently. This leads to the interactions of people with these robots evolving from being instant or one-time interactions into long-term interactions and even relationships. This paradigm shift in the social interactions of humans and robots sets new challenges to the field of sHRI. Not many approaches, techniques or implementations in sHRI were tested in such a long-term social interaction context, which is also true for the outlined pilot implementations of the SFAS framework in this thesis.

To address this need, both the scenario coverage (e.g. the hybrid use of SFAS along side natural language in a wider number of contextual scenarios), the scope (e.g. including other parameters of the voice signal than spectral shift in the voice alignment) and the experiment duration (e.g. the robots and the subjects interacting periodically for a longer duration of time) of the pilot implementations can be extended further.

For example, as (Zaga et al., 2016) suggested, designing robot behaviors with SFUs investigations should divert from purely focusing on the recognition of the behaviors or intentions, to also exploring how the recognition of the behaviors will affect the interaction. Such a diversion could also be explored in the context of long-term sHRI.

One of the first obvious questions for Semantic-Free Affective Speech regarding long-term sHRI is if the interaction partners will continue to perceive Gibberish speech positively despite the missing semantic. In instant interaction, not understanding a Gibberish speech, as long as it sounds as if it's a real language, still gives a natural feeling, which may be like getting exposed to a new foreign language for the first time. But it can be argued that when more interactions occur, people may have the expectation to start making sense of at least some of the utterances, especially for the ones that are repeated. As such, it can be speculated that repetitions of certain utterances in specific situational contexts may increase the feeling of naturalness, getting to know the robot better and attachment to the robot in long term sHRI for robots utilizing Semantic-Free Affective Speech.

As a first step a mechanism to achieve this can be implemented in SFAS framework, by tagging the spoken utterances with the situational context indexes in the EMOGIB database after their first use for the specific implementation. With such a mechanism, a robot for example would greet its interaction partner by vocalizing a speech including the same utterance tagged with the greeting context index every time. By time and experience through periodic interactions, the interaction partner may build a sense of vocabulary and grammar for the robot's specific language.

Some of this is of course speculative, but if these emerging predictions become a reality a bright future may exist for the extensive use of Semantic-Free Affective Speech in long-term social human robot interactions.

Bibliography

Abelin, Å., & Allwood, J. (2000). Cross linguistic interpretation of emotional prosody. In *Isca tutorial and research workshop (itrw) on speech and emotion* (p. 110-113).

Alty, J. L., Rigas, D., & Vickers, P. (2005). Music and speech in auditory interfaces: When is one mode more appropriate than another? In *International conference on auditory display* (p. 351-357).

Ambrus, D. C. (2000). *Collecting and recording of an emotional speech database* (Tech. Rep.). Faculty of Electrical Engineering, Institute of Electronics, Univ. of Maribor.

Annosoft Lipsync Tool 4.1. (n.d.). Retrieved from `http://www.annosoft.com/lipsync-tool` (Last accessed: March 2014)

Arslan, L. M., & Talkin, D. (1998). Speaker transformation using sentence HMM based alignments and detailed prosody modification. In *Proceedings of the ieee international conference on acoustics, speech and signal processing.* (Vol. 1, pp. 289–292).

Ayesh, A. (2006, 6-8 September). Structured sound based language for emotional robotic communicative interaction. In *The 15th ieee international symposium on robot and human interactive communication (roman 2006)* (pp. 135–140).

Ayesh, A. (2009, 1). Emotionally expressive music based interaction language for social robots. *ICGST International Journal on Automation, Robotics and Autonomous Systems*, *9*(1), 1 -10.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614–636.

Bänziger, T., Mortillaro, M., & Scherer, K. R. (2012). Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. *Emotion*, *12*(5), 1161.

Bänziger, T., & Scherer, K. R. (2010). Introducing the geneva multimodal emotion

portrayal (gemep) corpus. In K. R. Scherer, T. Bänziger, & E. Roesch (Eds.), *A blueprint for affective computing: A sourcebook and manual* (pp. 271–294). Oxford University Press Oxford, UK.

Barker, J., & Cooke, M. (1999). Is the sine-wave speech cocktail party worth attending? *Speech Communication*, *27*(3), 159–174.

Bartneck, C., & Michio, O. (2001). eMuu-An emotional robot. In *Proceedings of 2001 robofesta.* Citeseer.

Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russell, M. J., & Wong, M. (2004, May 26-28). " You Stupid Tin Box"-Children Interacting with the AIBO Robot: A Cross-linguistic Emotional Speech Corpus. In *Proceedings of the fourth international conference on language resources and evaluation (lrec).* Lisbon, Portugal.

Batliner, A., Hacker, C., Steidl, S., Nöth, E., & Haas, J. (2003). User states, user strategies, and system performance: how to match the one with the other. In *Isca tutorial and research workshop on error handling in spoken dialogue systems.*

Beck, A., Cañamero, L., Hiolle, A., Damiano, L., Cosi, P., Tesser, F., & Sommavilla, G. (2013). Interpretation of emotional body language displayed by a humanoid robot: A case study with children. *International Journal of Social Robotics*, *5*(3), 325-334.

Becker-Asano, C., & Ishiguro, H. (2009). Laughter in social robotics-no laughing matter. In *International workshop on social intelligence design (sid2009)* (pp. 287–300).

Becker-Asano, C., Kanda, T., Ishi, C., & Ishiguro, H. (2011). Studying laughter in combination with two humanoid robots. *AI & Society*, *26*(3), 291-300.

Belpaeme, T., Baxter, P., Greeff, J., Kennedy, J., Read, R., Looije, R., ... Zelati, M. C. (2013). Child-robot interaction: Perspectives and challenges. In G. Herrmann, J. Pearson Martin, A. Lenz, P. Bremner, A. Spiers, & U. Leonards (Eds.), *Social robotics* (Vol. 8239, p. 452-459). Springer International Publishing. doi: 10.1007/978-3-319-02675-6_45

Belpaeme, T., Baxter, P., Read, R., Wood, R., Cuayáhuitl, H., Kiefer, B., ... Humbert, R. (2012). Multimodal child-robot interaction: Building social bonds. *Journal of Human-Robot Interaction*, *1*(2), 33–53. doi: 10.5898/JHRI.1.2.Belpaeme

Bennett, K. (2004). *Linguistic steganography: Survey, analysis, and robustness*

concerns for hiding information in text (Tech. Rep.). Purdue University.

Blattner, M. M., Sumikawa, D. A., & Greenberg, R. M. (1989). Earcons and icons: Their structure and common design principles. *Human-Computer Interaction*, *4*(1), 11–44.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*, 49–59.

Bramas, B., Kim, Y.-M., & Kwon, D.-S. (2008). Design of a sound system to increase emotional expression impact in human-robot interaction. In *International conference on control, automation and systems (iccas 2008)* (pp. 2732–2737).

Breazeal, C. (2000). *Sociable machines: Expressive social exchange between humans and robots* (Unpublished doctoral dissertation). Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science.

Breazeal, C. (2002). *Designing sociable robots.* Cambridge, MA, USA: The MIT Press.

Breazeal, C. (2004). Social interactions in HRI: The robot view. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, *34*(2), 181–186.

Breazeal, C., & Aryananda, L. (2002). Recognition of affective communicative intent in robot-directed speech. *Autonomous robots*, *12*(1), 83–104.

Breazeal, C., Depalma, N., Orkin, J., & Chernova, S. (2013). Crowdsourcing human-robot interaction : New methods and system evaluation in a public environment. *Journal of Human-Robot Interaction*, *2*(1), 82–111. doi: 10.5898/JHRI.2.1.Breazeal

Broekens, J., & Brinkman, W.-P. (2013). AffectButton: a method for reliable and valid affective self-report. *International Journal of Human-Computer Studies*, *71*(6), 641 - 667.

Brooks, A., & Arkin, R. (2007). Behavioral overlays for non-verbal communication expression on a humanoid robot. *Autonomous Robots*, *22*(1), 55-74.

Brown, S. (2000). The "musilanguage" model of music evolution. In S. B. Nils L. Wallin Bjorn Merker (Ed.), (p. 271-300). MIT Press.

Burkhardt, F., Audibert, N., Malatesta, L., Türk, O., Arslan, L., & Auberge, V. (2006). Emotional prosody-does culture make a difference. In *Speech prosody*

(Vol. 2).

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. In *Interspeech* (Vol. 5, pp. 1517–1520).

Burkhardt, F., & Sendlmeier, W. (2000). Verification of acoustical correlates of emotional speech using formant-synthesis. In *Isca tutorial and research workshop (itrw) on speech and emotion* (pp. 151–156). Citeseer.

Busso, C., & Narayanan, S. S. (2008, May). Recording audio-visual emotional databases from actors: A closer look. In *Proceedings of the international conference on language resources and evaluation (lrec)* (p. 17-22). Marrakech, Morocco.

Cahn, J. (1990). The generation of affect in synthesized speech. *Journal of the American Voice I/O Society*, *8*, 1–19.

Chao, C., & Thomaz, A. (2013). Controlling social dynamics with a parametrized model of floor regulation. *Journal of Human-Robot Interaction*, *2*(1), 4 – 29. doi: 10.5898/JHRI.2.1.Chao

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, *2*, 113–124.

Connell, J. H. (2014). Extensible Grounding of Speech for Robot Instruction. In J. Markowitz (Ed.), (pp. 175–199). Berlin/Boston/Munich: De Gruyter.

Corveleyn, S., Coose, B., & Verhelst, W. (2002). Voice Modification and Conversion Using PLAR-Parameters. In *Ieee benelux workshop on model based processing and coding of audio (mpca).*

Cowie, R., & Cornelius, R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, *40*(1-2), 5–32.

Cowie, R., Douglas-Cowie, E., & Cox, C. (2005). Beyond emotion archetypes: Databases for emotion modelling using neural networks. *Neural networks*, *18*(4), 371–388.

Dall, R., Yamagishi, J., & King, S. (2014). Rating naturalness in speech synthesis: The effect of style and expectation. *Proceedings Speech Prosody, Dublin, Ireland.*

Deits, R., Tellex, S., Thaker, P., Simeonov, D., Kollar, T., & Roy, N. (2013). Clarifying commands with information-theoretic human-robot dialog. *Journal of Human-Robot Interaction*, *2*(2), 58–79.

Delaunay, F., de Greeff, J., & Belpaeme, T. (2010). A study of a retro-projected robotic face and its effectiveness for gaze reading by humans. In *Proceedings of the 5th international conference on human-robot interaction (hri'10)* (pp. 39 – 44). Osaka, Japan: ACM/IEEE.

Dingler, T., Lindsay, J., & Walker, B. N. (2008). Learnability of sound cues for environmental features: Auditory icons, earcons, spearcons, and speech. In *Proceedings of the 14th international conference on auditory display, paris, france* (pp. 1–6).

D'Mello, S., McCauley, L., & Markham, J. (2005). A mechanism for human-robot interaction through informal voice commands. In *Robot and human interactive communication, 2005. roman 2005. ieee international workshop on* (pp. 184–189).

Dombois, F., & Eckel, G. (2011). Audification. In T. Hermann, A. Hunt, & J. G. Neuhoff (Eds.), (pp. 301–324). Berlin, Germany: Logos Publishing House.

Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech communication*, *40*(1), 33–60.

Douglas-Cowie, E., Cowie, R., & Schröder, M. (2000). A new emotion database: considerations, sources and scope. In *Isca tutorial and research workshop (itrw) on speech and emotion.*

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., Mcrorie, M., ... others (2007). The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data. *Affective computing and intelligent interaction*, 488–500.

Dutoit, T., & Leich, H. (1993). MBR-PSOLA: Text-To-Speech synthesis based on an {MBE} re-synthesis of the segments database. *Speech Communication*, *13*(3-4), 435 - 440.

Ekman, P., & Friesen, W. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124–129.

Embgen, S., Luber, M., Becker-Asano, C., Ragni, M., Evers, V., & Arras, K. (2012). Robot-specific social cues in emotional body language. In *Proceedings of the 21st international symposium on robot and human interactive communication (ro-man 2012)* (pp. 1019–1025). Paris, France: IEEE. doi: 10.1109/ROMAN.2012.6343883

Esnaola, U., & Smithers, T. (2005, June). MiReLa: A musical robot. In *Proceedings*

of ieee international symposium on computational intelligence in robotics and automation (cira 2005) (pp. 67–72).

ETRO Audio-Visual Lab. (n.d.). Retrieved from `http://www.etrovub.be/Research/Nosey_Elephant_Studios/` (Last accessed: 25 May 2013)

Fridin, M., & Belokopytov, M. (2014). Embodied robot versus virtual agent: Involvement of preschool children in motor task performance. *International Journal of Human-Computer Interaction*, *30*(6), 459–469.

Friend, M. (2000). Developmental changes in sensitivity to vocal paralanguage. *Developmental Science*, *3*(2), 148–162.

Gabsdil, M. (2003). Clarification in spoken dialogue systems. In *Proceedings of the 2003 aaai spring symposium. workshop on natural language generation in spoken and written dialogue* (pp. 28–35).

Gardner, M. (1984). *Codes, ciphers and secret writing.* Mineola, N.Y: Dover Publications.

Gaut, B. (1993). The paradox of horror. *British Journal of Aesthetics*, *33*, 333–333.

Gaver, W. (1986). Auditory icons: Using sound in computer interfaces. *Human-Computer Interaction*, *2*(2), 167–177.

Gorostiza, J. F., & Salichs, M. A. (2011). End-user programming of a social robot by dialog. *Robotics and Autonomous Systems*, *59*(12), 1102–1114.

Grimm, M., Kroschel, K., & Narayanan, S. S. (2008, June). The Vera am Mittag German audio-visual emotional speech database. In *Proceedings of the ieee international conference on multimedia and expo (icme)* (p. 865-868). Hannover, Germany.

Hanich, J. (2009). Dis/liking disgust: the revulsion experience at the movies. *New Review of Film and Television Studies*, *7*(3), 293–309.

Hansen, J. H., Bou-Ghazale, S. E., Sarikaya, R., & Pellom, B. (1997). Getting started with SUSAS: a speech under simulated and actual stress database. In *Eurospeech* (Vol. 97, pp. 1743–46).

Haring, M., Bee, N., & André, E. (2011). Creation and evaluation of emotion expression with body movement, sound and eye color for humanoid robots. In *Ro-man, 2011 ieee* (pp. 204–209).

Hart, M. (1971). *Project Gutenberg.* Project Gutenberg. Retrieved from `http://www.gutenberg.org` (Last accessed: March 2014)

Hermann, T., Hunt, A., & Neuhoff, J. G. (Eds.). (2011). *The Sonification Handbook*. Berlin, Germany: Logos Verlag.

Hinton, L., Nichols, J., & Ohala, J. J. (1994). Sound symbolism. In J. Hinton, J. Nichols, & J. J. Ohala (Eds.), (pp. 325–47). Cambridge University Press.

Holzapfel, H., & Gieselmann, P. (2004). A way out of dead end situations in dialogue systems for human-robot interaction. In *Humanoid robots, 2004 4th ieee/ras international conference on* (Vol. 1, pp. 184–195).

Iida, A., & Campbell, N. (2003). Speech database design for a concatenative text-to-speech synthesis system for individuals with communication disorders. *International Journal of Speech Technology*, *6*(4), 379–392.

Imai, M., Hiraki, K., Miyasato, T., Nakatsu, R., & Anzai, Y. (2003). Interaction with robots: Physical constraints on the interpretation of demonstrative pronouns. *International Journal of Human-Computer Interaction*, *16*(2), 367–384.

Iriondo, I., Planet, S., Socoró, J.-C., & Alías, F. (2007). Objective and subjective evaluation of an expressive speech corpus. In *International conference on nonlinear speech processing* (pp. 86–94).

Jee, E., Jeong, Y., Kim, C., & Kobayashi, H. (2010). Sound design for emotion and intention expression of socially interactive robots. *Intelligent Service Robotics*, *3*, 199–206.

Jee, E.-S., Kim, C. H., Park, S.-Y., & Lee, K.-W. (2007, August). Composition of musical sound expressing an emotion of robot based on musical factors. In *Proceedings of the 16th international symposium on robot and human interactive communication (ro-man 2007)* (pp. 637–641). Jeju Island, Korea: IEEE.

Jee, E.-S., Park, S.-Y., Kim, C. H., & Kobayashi, H. (2009, September). Composition of musical sound to express robot's emotion with intensity and synchronized expression with robot's behavior. In *Proceedings of the 18th international symposium on robot and human interactive communication (ro-man 2009)* (pp. 369–374). Toyama, Japan: IEEE.

Johannsen, G. (2001). Auditory displays in human-machine interfaces of mobile robots for non-linguistic speech communication with humans. *Journal of Intelligent and Robotic Systems*, *32*(2), 161-169.

Johannsen, G. (2002). Auditory display of directions and states for mobile systems. In *Proceedings of the international conference on auditory display* (pp. 98–103).

Johannsen, G. (2004, Apr). Auditory displays in human-machine interfaces. *Proceedings of the IEEE*, *92*(4), 742-758.

Johnson, W. F., Emde, R. N., Scherer, K. R., & Klinnert, M. D. (1986). Recognition of emotion from vocal cues. *Archives of General Psychiatry*, *43*(3), 280-283.

Jung, H., Seon, C.-N., Kim, J. H., Sohn, J. C., Sung, W.-K., & Park, D.-I. (2005). Information extraction for users utterance processing on ubiquitous robot companion. In *Natural language processing and information systems* (pp. 337–340). Springer.

Juslin, P. N., & Laukka, P. (2003, September). Communication of emotions in vocal expression and music performance: different channels, same code? *Psychological bulletin*, *129*(5), 770–814.

Knoll, M., Uther, M., & Costall, A. (2009). Effects of low-pass filtering on the judgment of vocal affect in speech directed to infants, adults and foreigners. *Speech Communication*, *51*(3), 210 – 216. doi: 10.1016/j.specom.2008.08.001

Kobayashi, T., & Fujie, S. (2013). Conversational robots: An approach to conversation protocol issues that utilizes the paralinguistic information available in a robot-human setting. *Acoustical Science and Technology*, *34*(2), 64-72.

Komatsu, T. (2005). Toward making humans empathize with artificial agents by means of subtle expressions. In *1st international conference on affective computing and intelligent interaction (acii2005)* (pp. 458 – 465). Bejing, China.

Komatsu, T., & Kobayashi, K. (2012). Can users live with overconfident or unconfident systems?: A comparison of artificial subtle expressions with human-like expression. In *Proceedings of conference on human factors in computing systems (chi 2012)* (pp. 1595–1600). Austin, Texas.

Komatsu, T., & Yamada, S. (2007). How appearance of robotic agents affects how people interpret the agents' attitudes. In *Proceedings of the international conference on advances in computer entertainment technology - ace '07* (pp. 123–126). New York, NY, USA: ACM Press. doi: 10.1145/1255047.1255071

Komatsu, T., & Yamada, S. (2008). How does appearance of agents affect how people interpret the agents' attitudes: An Experimental investigation on expressing the same information from agents having different appearance. In *Ieee congress on evolutionary computation* (pp. 1935–1940).

Komatsu, T., & Yamada, S. (2011, February). How does the agents' appearance affect users' interpretation of the agents' attitudes: Experimental investigation on expressing the same artificial sounds from agents with different appear-

ances. *International Journal of Human-Computer Interaction*, *27*(3), 260–279.

Komatsu, T., Yamada, S., Kobayashi, K., Funakoshi, K., & Nakano, M. (2010). Artificial subtle expressions: Intuitive notification methodology of artifacts. In *Proceedings of the 28th international conference on human factors in computing systems (chi'10)* (pp. 1941–1944). New York, New York, USA: ACM.

Kozima, H., Michalowski, M. P., & Nakagawa, C. (2009, November). Keepon: A playful robot for research, therapy, and entertainment. *International Journal of Social Robotics*, *1*(1), 3–18.

Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J., ... Conaway, M. (2002). The impact of "No Opinion" response options on data quality: Non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly*, *66*(3), 371-403.

Latacz, L., Kong, Y., Mattheyses, W., & Verhelst, W. (2008, 9). An overview of the VUB entry for the 2008 blizzard challenge. In *Proceedings of the interspeech blizzard challenge.* Brisbane, Australia.

Laukka, P. (2005, September). Categorical perception of vocal emotion expressions. *Emotion (Washington, D.C.)*, *5*(3), 277–95.

Leite, I., Pereira, A., Martinho, C., & Paiva, A. (2008). Are emotional robots more fun to play with? In *Robot and human interactive communication, 2008. ro-man 2008. the 17th ieee international symposium on* (pp. 77–82).

Liberman, M., Davis, K., Grossman, M., Martey, N., & Bell, J. (2002). Emotional Prosody Speech and Transcripts. In *Linguistic data consortium.* Philadelphia.

Libin, A. V., & Libin, E. V. (2004, Nov). Person-robot interactions from the robopsychologists' point of view: the robotic psychology and robotherapy approach. *Proceedings of the IEEE*, *92*(11), 1789–1803. doi: 10.1109/JPROC.2004.835366

Lison, P., & Kruiff, G.-J. (2009, September). Robust processing of situated spoken dialogue. In *Proceedings of the 32nd annual german conference on advances in artificial intelligence (ki'09)* (pp. 241–248). Paderbron, Germany: Springer-Verlag.

Lohse, M., Rohlfing, K. J., Wrede, B., & Sagerer, G. (2008, May). "Try something else!", When users change their discursive behaviour in human-robot interaction. In *Proceedings of the international conference on robotics and automation (icra 2008)* (pp. 3481–3486). Pasadena, California, U.S.A.: IEEE.

Mac, D.-K., Aubergé, V., Rilliard, A., & Castelli, E. (2010). Cross-cultural perception of Vietnamese Audio-Visual prosodic attitudes. In *Speech prosody 2010-fifth international conference.*

Mattheyses, W. (2013). *A multimodal approach to audiovisual textto-speech synthesis* (Unpublished doctoral dissertation). Ph. D. thesis, Vrije Universiteit Brussel.

McCartney, J. (2002). Rethinking the computer music language: SuperCollider. *Computer Music Journal*, *26*(4), 61– 68.

Mitchell, W. J., Szerszen Sr, K. A., Lu, A. S., Schermerhorn, P. W., Scheutz, M., & MacDorman, K. F. (2011). A mismatch in the human realism of face and voice produces an uncanny valley. *i-Perception*, *2*(1), 10–12.

Moore, K., Roger. (2014). Spoken language processing: Time to look outside? In L. Besacier, A.-H. Dediu, & C. Martín-Vide (Eds.), *Statistical language and speech processing* (p. 21-36). Springer International Publishing.

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

Mower, E., Mataric, M. J., & Narayanan, S. (2009). Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information. *IEEE Transactions on Multimedia*, *11*(5), 843–855.

Mozos, O. M., Jensfelt, P., Zender, H., Kruijff, G.-J. M., & Burgard, W. (2007). From labels to semantics: An integrated system for conceptual spatial representations of indoor environments for mobile robots. In *Proceedings of the ieee icra workshop: Semantic information in robotics* (pp. 33–40).

Mubin, O., Bartneck, C., & Feijs, L. (2009). What you say is not what you get: Arguing for artificial languages instead of natural languages in human robot speech interaction. In *the spoken dialogue and human-robot interaction workshop at ieee roman 2009.* Toyama, Japan.

Mubin, O., Bartneck, C., & Feijs, L. (2010). Towards the design and evaluation of ROILA: a speech recognition friendly artificial language. In *Proceedings of the 7th international conference on advances in natural language processing (icetal'10)* (pp. 250–256). Reykjavik, Iceland: Springer-Verlag.

Mubin, O., Bartneck, C., Feijs, L., Hooft van Huysduynen, H., Hu, J., & Muelver, J. (2012). Improving speech recognition with the robot interaction language. *Disruptive Science and Technology*, *1*(2), 79–88.

Mumm, J., & Mutlu, B. (2011). Human-robot proxemics: Physical and psycholog-

ical distancing in human-robot interaction. In *Proceedings of the 6th international conference on human-robot interaction (hri'11)* (pp. 331–338). Lausanne, Switzerland.

Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, *93*(2), 1097–1108.

Murray, I. R., & Arnott, J. L. (1995). Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Communication*, *16*(4), 369–390.

Murray, I. R., & Arnott, J. L. (1996). Synthesizing emotions in speech: is it time to get excited? In *Proceeding of fourth international conference on spoken language processing (icslp 96)* (pp. 1816–1819). IEEE.

NAO 2015 - SoftBank Robotics' humaonid robotic platform. (n.d.). Retrieved from `https://www.ald.softbankrobotics.com/en/cool-robots/nao` (Last accessed: 11 Jul 2017)

Neill, A., & Ridley, A. (2013). *Arguing about art: contemporary philosophical debates.* Routledge.

Németh, G., Olaszy, G., & Csapó, T. G. (2011). Spemoticons: Text-to-speech based emotional auditory cues. In *The 17th international conference on auditory display* (pp. 1–7).

Niewiadomski, R., Hofmann, J., Urbain, J., Platt, T., Wagner, J., Piot, B., ... others (2013). Laugh-aware virtual agent and its impact on user amusement. In *Proceedings of the 2013 international conference on autonomous agents and multi-agent systems* (pp. 619–626).

Olaszy, G., Németh, G., Olaszi, P., Kiss, G., Zainkó, C., & Gordos, G. (2000). Profivox - A hungarian text-to-speech system for telecommunications applications. *International Journal of Speech Technology*, *3*(3-4), 201-215.

Olive, J., & Buchsbaum, A. (1987, July). *Changing voice characteristics in text to speech synthesis* (Technical Memorandum). AT&T Bell-Labs.

Oudeyer, P.-Y. (2003). The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies*, *59*(1-2), 157–183. doi: http://dx.doi.org/10.1016/S1071-5819(02)00141-6

Palladino, D. K., & Walker, B. N. (2007). Learning rates for auditory menus enhanced with spearcons versus earcons. In *Proceedings of the 13th international*

conference on auditory display (pp. 274–279).

Paulmann, S., & Pell, M. (2011). Is there an advantage for recognizing multi-modal emotional stimuli? *Motivation and Emotion*, *35*(2), 192-201.

Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S. A. (2009). Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics*, *37*(4), 417 - 435.

Picard, R. W. (1997). *Affective Computing.* Cambridge, MA, U.S.A.: MIT Press.

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, *17*(3), 715–734.

Prendinger, H., Becker, C., & Ishizuka, M. (2006). A study in users' physiological response to an empathic interface agent. *International Journal of Humanoid Robotics*, *3*, 371–391.

Rae, I., Takayama, L., & Mutlu, B. (2013, March). The influence of height in robot-mediated communication. In *Proceedings of the 8th international conference on human-robot interaction (hri'13)* (pp. 1–8). Tokyo, Japan: IEEE. doi: 10.1109/HRI.2013.6483495

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. *The Quarterly Journal of Experimental Psychology Section A*, *55*(4), 1339-1362.

Read, R., & Belpaeme, T. (2010, October). Interpreting non-linguistic utterances by robots : Studying the influence of physical appearance. In *Proceedings of the 3rd international workshop on affective interaction in natural environments (affine 2010) at acm multimedia 2010* (pp. 65–70). Firenze, Italy: ACM.

Read, R., & Belpaeme, T. (2012, March). How to use non-linguistic utterances to convey emotion in child-robot interaction. In *Proceedings of the 7th international conference on human-robot interaction (hri'12)* (pp. 219–220). Boston, MA, U.S.A.: ACM/IEEE.

Read, R., & Belpaeme, T. (2013, March). People interpret robotic non-linguistic utterances categorically. In *Proceedings of the 8th international conference on human-robot interaction (hri'13)* (pp. 209–210). Tokyo, Japan: ACM/IEEE.

Read, R., & Belpaeme, T. (2014a, March). Non-Linguistic Utterances Should be Used Alongside Language, Rather than on their Own or as a Replacement. In *Proceedings of the 9th international conference on human-robot interaction*

(hri14).

Read, R., & Belpaeme, T. (2014b). Situational context directs how people affectively interpret robotic non-linguistic utterances. In *Proceedings of the 9th international conference on human-robot interaction (hri'14)* (pp. 41 – 48). Bielefeld, Germany: ACM/IEEE.

Remez, R., & Rubin, P. (1993). On the intonation of sinusoidal sentences: Contour and pitch height. *The Journal of the Acoustical Society of America*, *94*(4), 1983-1988.

Remez, R., Rubin, P., Pisoni, D., & Carrell, T. (1981). Speech perception without traditional speech cues. *Science*, *212*, 947–950.

Ribeiro, T., & Paiva, A. (2012). The Illusion of Robotic Life. In *Proceedings of the 7th international conference on human-robot interaction (hri'12)* (pp. 383–390). Boston, MA.

Ros Espinoza, R., Nalin, M., Wood, R., Baxter, P., Looije, R., Demiris, Y., & Belpaeme, T. (2011, November). Child-robot interaction in the wild: Advice to the aspiring experimenter. In *Proceedings of the 13th international conference on multimodal interfaces (icmi'11)* (pp. 335–342). Valencia, Spain: ACM.

Saerbeck, M., & Bartneck, C. (2010). Perception of affect elicited by robot motion. In *Proceedings of the 5th international conference on human-robot interaction (hri'10)* (pp. 53–60). Osaka, Japan: ACM/IEEE.

Saerbeck, M., Schut, T., Bartneck, C., & Janse, M. D. (2010). Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1613–1622).

Saldien, J. (2009). *Development of the huggable social robot Probo: on the conceptual design and software architecture* (Unpublished doctoral dissertation). Vrije Universiteit Brussel. Faculty of Engineering.

Saldien, J., Goris, K., Vanderborght, B., Vanderfaeillie, J., & Lefeber, D. (2010). Expressing Emotions with the Social Robot Probo. *International Journal of Social Robotics*, *2*(4), 377–389. doi: 10.1007/s12369-010-0067-6

Saldien, J., Goris, K., Yilmazyildiz, S., Verhelst, W., & Lefeber, D. (2008). On the design of the huggable robot Probo. *Journal of Physical Agents*, *2*(2), 3-11. doi: 10.14198/JoPha.2008.2.2.02

Saratxaga, I., Navas, E., Hernáez, I., & Luengo, I. (2006). Designing and recording

an emotional speech database for corpus based synthesis in Basque. In *Proc. of fifth international conference on language resources and evaluation (lrec)* (pp. 2126–2129).

Scherer, K. (1971). Randomized splicing: a note on a simple technique for masking speech content. *Journal of Experimental Research in Personality*, *5*, 155–159.

Scherer, K. (1985). Vocal affect signalling: A comparative approach. In J. Rosenblatt, C. Beer, M.-C. Busnel, & P. Slater (Eds.), *Advances in the study of behavior* (pp. 189–244). New York, USA: Academic Press.

Scherer, K. (1986, March). Vocal affect expression: a review and a model for future research. *Psychological Bulletin*, *99*(2), 143–65.

Scherer, K. (1995, October). Expression of emotion in voice and music. *Journal of Voice*, *9*(3), 235–248.

Scherer, K. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1-2), 227–256.

Scherer, K., & Ekman, P. (1982). Methods of research on vocal communication: Paradigms and parameters. In K. Scherer & P. Ekman (Eds.), *Handbook of methods in nonverbal behavior research* (pp. 136–198). Cambridge, UK: Cambridge University Press.

Scherer, K., Koivumaki, J., & Rosenthal, R. (1972). Minimal cues in the vocal communication of affect: Judging emotions from content-masked speech. *Journal of Psycholinguistic Research*, *1*(3), 269–285.

Scherer, K., & Oshinsky, J. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion*, *1*(4), 331–346.

Scherer, K. R. (1994). Affect bursts. In S. van Goozen, N. E. van de Poll, & J. A. Sergeant (Eds.), *Emotions: Essays on emotion theory* (S. van Goozen, N.E. van de Poll, & J.A. Sergeant (Eds.) ed., p. 161-196). Hillsdale, USA: NJ: Erlbaum.

Schröder, M. (2001). Emotional speech synthesis: A review. In *Proceedings of the 7th european conference on speech communication and technology (eurospeech 2001)* (pp. 2–5). Aalborg, Denmark.

Schröder, M. (2003a). Experimental study of affect bursts. *Speech Communication*, *40*(1-2), 99–116.

Schröder, M. (2003b). *Speech and emotion research: An overview of research*

frameworks and a dimensional approach to emotional speech synthesis (Unpublished doctoral dissertation). Institute of Phonetics, Saarland University, Saarbrücken, Germany.

Schröder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., ... Wollmer, M. (2012). Building autonomous sensitive artificial listeners. *Transactions on Affective Computing*, *3*(2), 165–183. doi: 10.1109/T-AFFC.2011.34

Schröder, M., Burkhardt, F., & Krstulovic, S. (2010). Synthesis of emotional speech. In K. R. Scherer, T. Bänziger, & E. Roesch (Eds.), *Blueprint for affective computing* (pp. 222 – 231). Oxford, UK: Oxford University Press.

Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., & Gielen, S. (2001, 9). Acoustic correlates of emotion dimensions in view of speech synthesis. In *Eurospeech 2001* (pp. 87–90).

Schröder, M., & Trouvain, J. (2003). The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *International Journal of Speech Technology*, *6*(4), 365-377. doi: 10.1023/A:1025708916924

Schuller, B., & Batliner, A. (2014). *Computational paralinguistics: Emotion, affect and personality in speech and language processing.* John Wiley & Sons.

Schwent, M., & Arras, K. (2014). R2-D2 reloaded: A flexible sound synthesis system for sonic human-robot interaction design. In *Proceedings of the 23rd international symposium on robot and human interaction communiation (roman 2014)* (p. 161 - 167). Edinburgh, UK.

Seo, S. H., Geiskkovitch, D., Nakane, M., King, C., & Young, J. E. (2015). Poor Thing! Would You Feel Sorry for a Simulated Robot?: A comparison of empathy toward a physical and a simulated robot. In *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction* (pp. 125–132).

Shiwa, T., Kanda, T., Imai, M., Ishiguro, H., & Hagita, N. (2009, February). How quickly should a communication robot respond? Delaying strategies and habituation effects. *International Journal of Social Robotics*, *1*(2), 141–155.

Shochi, T., Aubergé, V., & Rilliard, A. (2006). How prosodic attitudes can be false friends: Japanese vs. French social affects. *Speech Prosody, Dresden*, 692–696.

Silva-Pereyra, J., Conboy, B. T., Klarman, L., & Kuhl, P. K. (2007). Grammatical processing without semantics? An event-related brain potential study of preschoolers using jabberwocky sentences. *Journal of cognitive neuroscience*, *19*(6), 1050–1065.

Singh, A., & Young, J. (2012). Animal-inspired human-robot interaction: A robotic tail for communicating state. In *Proceedings of the 7th international conference on human-robot interactinon (hri'12)* (pp. 237–238). Boston, USA.

Snel, J., & Cullen, C. (2013). Judging emotion from low-pass filtered naturalistic emotional speech. In *Affective computing and intelligent interaction (acii), 2013 humaine association conference on* (pp. 336–342).

Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Science*, *7*(7), 308 – 312.

Steels, L., Kaplan, F., McIntyre, A., & Van Looveren, J. (2002). Crucial factors in the origins of word-meaning. In A. Wray (Ed.), *The transition to language* (pp. 252–271). Oxford: Oxford University Press.

Takayama, L., & Pantofaru, C. (2009, Oct). Influences on proxemic behaviors in human-robot interaction. In *Proceedings of the international conference on intelligent robots and systems (iros'09)* (p. 5495-5502). St. Louis, USA.

Teshigawara, M., Amir, N., Amir, O., Wlosko, E. M., & Avivi, M. (2007). Effects of random splicing on listeners perceptions. In J. Trouvain & W. J. Barry (Eds.), *16th international congress of phonetic sciences (icphs)* (p. 2101 - 2104).

Theobalt, C., Bos, J., Chapman, T., Espinosa-Romero, A., Fraser, M., Hayes, G., ... Reeve, R. (2002). Talking to Godot: Dialogue with a mobile robot. In *Ieee/rsj international conference on intelligent robots and systems, 2002.* (Vol. 2, pp. 1338–1343). doi: 10.1109/IRDS.2002.1043940

Tickle, A. (2000). English and Japanese speaker's emotion vocalizations and recognition: a comparison highlighting vowel quality. In *Isca workshop on speech and emotion* (p. 157-183).

Trouvain, J., & Schröder, M. (2004). How (not) to add laughter to synthetic speech. In E. André, L. Dybkjaer, W. Minker, & P. Heisterkamp (Eds.), *Affective dialogue systems* (Vol. 3068, p. 229-232). Springer Berlin Heidelberg.

Trovato, G., Zecca, M., Kishi, T., Endo, N., Hashimoto, K., & Takanishi, A. (2013). Generation of humanoid robot's facial expressions for context-aware communication. *International Journal of Humanoid Robotics*, *10*, 1350013-1 – 1350013-23.

Van Tassel, D. (1969). Cryptographic techniques for computers. In *Proceedings of the may 14-16, 1969, spring joint computer conference* (pp. 367–372).

Vatsa, A., Mohan, T., & Vatsa, S. (2012). Novel cipher technique using substitution

method. *International Journal of Information and Network Security (IJINS)*, *1*(4), 313-320.

Vazquez, M., Steinfeld, A., Hudson, S. E., & Forlizzi, J. (2014). Spatial and other social engagement cues in a child-robot interaction: effects of a sidekick. In *Proceedings of the 9th international conference on human-robot interaction (hri'14)* (p. 391 - 398). Bielefeld, Germany: ACM/IEEE.

Verhelst, W., & Roelands, M. (1993). An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech. In *Ieee international conference on acoustics, speech, and signal processing (icassp)* (Vol. 2, pp. 554–557).

Ververidis, D., & Kotropoulos, C. (2003). A State of the Art Review on Emotional Speech Databases. In *1st richmedia conference proceedings* (p. 109 - 119). Lausanne, Switzerland.

Vickers, P., & Alty, J. L. (2002). Using music to communicate computing information. *Interacting with Computers*, *14*(5), 435–456.

Walker, B. N., Nance, A., & Lindsay, J. (2006). Spearcons: Speech-based earcons improve navigation performance in auditory menus. In *Proceedings of the international conference on auditory display, london, uk* (pp. 63–68).

Walters, M. L., Syrdal, D. S., Dautenhahn, K., te Boekhorst, R., & Koay, K. L. (2007). Avoiding the uncanny valley: robot appearance, personality and consistency of behaviour in an attention-seeking home scenario for a robot companion. *Autonomous Robots*, *24*(2), 159–178.

Wang, W., Athanasopoulos, G., Yilmazyildiz, S., Patsis, G., Enescu, V., Sahli, H., ... Canamero, L. (2014, September). Natural emotion elicitation for emotion modeling in child-robot interactions. In *4th workshop on child-computer interaction (wocci 2014) at interspeech.* Singapore, Singapore.

Ward, N. (1996). Using prosodic clues to decide when to produce back-channel utterances. In *Spoken language, 1996. icslp 96. proceedings., fourth international conference on* (Vol. 3, pp. 1728–1731).

Wilson, B. J., & Weiss, A. J. (1993). The effects of sibling coviewing on preschoolers' reactions to a suspenseful movie scene. *Communication Research*, *20*(2), 214–248.

Yang, P.-F., & Stylianou, Y. (1998). Real Time Voice Alteration Based on Linear Prediction. In *Proceedings of icslp* (pp. 1667–1670). Sydney, Australia.

Yilmazyildiz, S. (2006). *Communication of emotions for e-creatures* (Unpublished master's thesis). Vrije Universiteit Brussel.

Yilmazyildiz, S., Athanasopoulos, G., Patsis, G., Wang, W., Oveneke, M. C., Latacz, L., ... Lefeber, D. (2013). Voice modification for wizard-of-oz experiments in robot-child interaction. In *Proceedings of the workshop on affective social speech signals.* Grenoble, France.

Yilmazyildiz, S., Henderickx, D., Vanderborght, B., Verhelst, W., Soetens, E., & Lefeber, D. (2011, October). EMOGIB : Emotional gibberish speech database for affective human-robot interaction. In *Proceedings of the international conference on affective computing and intelligent interaction (acii'11)* (pp. 163–172). Memphis, TN, U.S.A.: Springer-Verlag.

Yilmazyildiz, S., Henderickx, D., Vanderborght, B., Verhelst, W., Soetens, E., & Lefeber, D. (2013). Multi-modal emotion expression for affective human-robot interaction. In *Proceedings of the workshop on affective social speech signals (wasss 2013).* Grenoble, France.

Yilmazyildiz, S., Latacz, L., Mattheyses, W., & Verhelst, W. (2010, September). Expressive gibberish speech synthesis for affective human-computer interaction. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Text, speech and dialogue: 13th international conference, tsd 2010, brno, czech republic, september 6-10, 2010. proceedings* (pp. 584–590). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-642-15760-8_74

Yilmazyildiz, S., Mattheyses, W., Patsis, Y., & Verhelst, W. (2006). Expressive speech recognition and synthesis as enabling technologies for affective robot-child communication. *Advances in Multimedia Information Processing - PCM 2006, Lecture Notes in Computer Science (LNCS), 4261*, 1–8.

Yilmazyildiz, S., Patsis, G., Verhelst, W., Henderickx, D., Soetens, E., Athanasopoulos, G., ... Lefeber, D. (2012). Voice Style Study for Human-Friendly Robots: Influence of the Physical Appearance. In *5th international workshop on human-friendly robotics (hfr2012).*

Yilmazyildiz, S., Read, R., Belpeame, T., & Verhelst, W. (2016). Review of Semantic-Free Utterances in Social Human-Robot Interaction. *International Journal of Human-Computer Interaction, 32*(1), 63–85. doi: 10.1080/10447318.2015.1093856

Yilmazyildiz, S., Verhelst, W., & Sahli, H. (2015, Nov). Gibberish speech as a tool for the study of affective expressiveness for robotic agents. *Multimedia Tools and Applications, 74*(22), 9959–9982. doi: 10.1007/s11042-014-2165-1

Zaga, C., Vries, R. A. D., Spenkelink, S. J., Truong, K. P., & Evers, V. (2016). Help-Giving Robot Behaviors in Child-Robot Games: Exploring Semantic Free Utterances. In *11th acm/iee international conference of human robot interaction.* Christchurch, New Zealand.