MoSS: Unfolding Playful Imaginaries of Synthetic Voice Design through a Modular Smart Speaker

Søren LyngsøKnudsen
IT University of Copenhagen, Dept. of Digital Design
slkn@itu.dk

Jonas Fritsch
IT University of Copenhagen, Dept. of Digital Design
frit@itu.dk

Stina Hasse Jørgensen
IT University of Copenhagen, Dept. of Digital Design
shaj@itu.dk

ABSTRACT

In this paper we introduce the experimental prototype MoSS (Modular Smart Speaker). MoSS is a modular speech synthesis smart speaker system, and a feature-rich and capable platform that resembles and functions as an intelligent smart speaker, while featuring extended functions for advanced audio processing. MoSS is composed by a number of software and hardware components, allowing people to customize, modify and modulate its synthetic voice in real-time. Through playfully combining the different tools and parameters, MoSS facilitates an investigation of human/machine interaction with voice controlled smart home objects, with a focus on the exploration of sound and voice synthesis. We present the process leading the creation of MoSS, how it works and the preliminary results from using the device to inspire novel vocal imaginaries in synthetic voice design.

CCS CONCEPTS

• **Human-centered Computing**; • **Interaction Design**; • **Systems and Tools for Interaction Design**;

KEYWORDS

Smart Speakers, Synthetic Voice Design, Exploratory Prototyping, Modular Synthesizers

ACM Reference Format:
Søren LyngsøKnudsen, Jonas Fritsch, and Stina Hasse Jørgensen. 2024. MoSS: Unfolding Playful Imaginaries of Synthetic Voice Design through a Modular Smart Speaker. In *Designing Interactive Systems Conference (DIS Companion '24), July 01–05, 2024, IT University of Copenhagen, Denmark.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3656156.3665428

1 INTRODUCTION

Synthetic voices have been part of our everyday culture for over a decade, especially as the main point of interaction in smart home devices such as Google's Audio Nest, Apple's HomePod, and Amazon's Echo, in which they broadcast news, weather forecasts, reminders, messages and make calls on the request of its interlocutor. The devices and their synthetic voices have been influencing how we interact with and listen to different voice qualities and vocal expressions [8]. The design of the speech synthesis systems in the smart home devices often presents its users with specified vocal expressions and voice qualities developed based on company values for what creates good user interactions and experiences [1]. Currently, machine learning tools have rapidly changed the voice synthesis landscape, making it possible for more people to play with, create and modify synthetic voices in new ways, exploring the breath of vocal expressions [4]. It is therefore both easier and more important than ever to explore the potential of synthetic voice designs that move beyond existing technologies' ideals of intelligibility, seamless interaction, user friendliness and 'human-level quality' [5–7, 9].

In this paper, we present an exploratory prototype that allows for the modulation of synthetic voices in real-time to experiment with different forms of synthetic vocality. MoSS (Modular Smart Speaker) combines the interactive voice control of a smart speaker with the playful hands-on interaction of a modular synthesizer. The modularity of the hybrid, both in hardware and software, makes it possible to patch together a wide variety of speech synthesis and audio processing tools to explore new sonic territory in the context of speech synthesis. Rather than being a passive recipient of a set vocality, users can interact and explore the vocal expressions and voice qualities as an interactive, aesthetic and playable part of the smart speaker device; an engagement, which might lead to changes in our everyday sonic cultures on a broader scale. Based on our experience of using the MoSS prototype in a workshop setting with sound and design professionals, we argue that MoSS is a technology of play – a *plaything* [10] – that allows us to materialize alternative futures and imaginaries of living with diverse synthetic vocality.
We first situate our work in relation to existing smart speakers and their infrastructural components and design affordances. We then present the design process leading to the creation of MoSS, insights gained from this and the technical description of the final prototype. This is followed by preliminary results from a workshop using MoSS to activate novel vocal imaginaries. Finally, we discuss findings from these explorations and point to future.

2 RELATED WORK: SMART SPEAKER DESIGN

In this paper, we primarily relate to smart speakers as a specific area of application for synthetic voice design. Examples of the most popular smart speakers today include the Amazon Echo/Amazon Echo Dot, Sonos One, Apple HomePod Mini, Google Nest Audio, and JBL Link View. These devices typically have a microphone for voice commands, a built-in virtual assistant with a synthetic voice, and the ability to control smart home devices. There are clear similarities in appearance between most smart speakers on the market, with variations in shape and finish. These devices are covered in fabric and use LED lights for visual communication.

DIS Companion '24, July 01–05, 2024, IT University of Copenhagen, Denmark

ACM ISBN 979-8-4007-0632-5/24/07
https://doi.org/10.1145/3656156.3665428

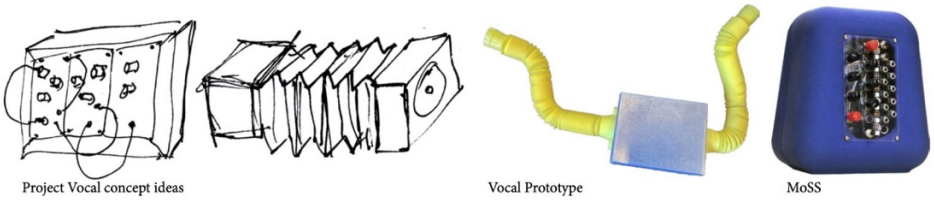

Figure 1: Different design proposals connected to the MoSS concept

Most smart speakers have similar control schemes, which usually consist of touch controls and mic mute switches.

To interact with a smart speaker, the device must be activated. The two most common ways to activate the speaker are either through wake word detection or by pressing a button. Wake word detection involves the device continuously recording audio and analyzing the stream for a pre-defined 'wake word' (e.g. 'Ok Google' or 'Hey Siri'). Either method puts the device into active listening mode, which prompts a second recording for the user to speak their commands. The recording is then automatically analyzed and transcribed using speech-to-text software (STT). This is usually done via cloud-based software although some devices also use built-in STT. The intelligent assistant software uses transcribed speech for its input, but the implementations vary greatly in both method and complexity. The majority, however, rely on Natural Language Processing and machine learning [3]. Most commercially available smart speakers are typically used for answering general knowledge questions, playing audio content and for home automation. For this type of smart speaker, its functionality framed as a personal or voice "assistant" is limited and further interaction with the user is minimal.

Although they share some design similarities, smart speakers differ in terms of sound quality and integration with other speaker devices. They are primarily designed to mimic human conversation in interaction with its users and thus reply with a humanlike voice [2]. The reply of smart speakers is initially generated as text and then transformed into speech through the process of text-to-speech (TTS). There are numerous types of TTS with different approaches to synthesizing natural-sounding speech. Depending on the type and complexity of the TTS software, it can either run locally on the device, or it can connect to cloud-based services. Smart speakers typically have limited computational power and are thus not able to perform heavy tasks e.g. generating natural sounding voices without the introduction of significant latency. Internet access makes it possible for them to use cloud-based services for TTS etc. by transmitting audio and other data to and from the device.

3 DESIGNING THE MOSS PROTOTYPE

The MoSS prototype was designed as part of an ongoing research project entitled *Voice as a Matter of Design: a Framework for Novel Vocal Imaginaries.* The project critically investigates the affective and sociocultural implications for human-machine configuration in a time where voices are increasingly designed. A major aim in the project is to explore and facilitate a broader spectrum of vocal expressions operating beyond normative vocal stereotypes towards developing a framework for pluralistic synthetic voice design. Beyond collective, situated listening experiments with different kinds of synthetic voices, we also wanted to create an infrastructure that would allow us and project participants to actively filter and modulate existing synthetic voices in real-time to open novel vocal imaginaries. In the following, we briefly outline the process that led to the creation of MoSS, before going more into detail with its core functionality.

When embarking on the design process, we first explored several prototype concepts for alternative smart speakers, each of them designed to demonstrate and explore a certain aspect of sound and interaction. The overall paradigm in our design proposals was to create filters for a smart speaker's sound and involved exploring spatialization and resonance, multichannel sound and choral qualities, extended sound processing and synthesis (Fig. 1). Whereas smart speakers generally can play audio content and perform home automation, we decided to only focus on the general knowledge features of smart speakers. What started out as a separate concept was later incorporated into the single design that became MoSS. Here, we present some of the main considerations for the selection and implementation of software and hardware for sound processing during prototyping.

A technical outset for the exploration was the Eurorack modular synthesizer that provides musicians with a versatile platform for sound design and experimentation, offering customization, flexibility, and creative possibilities through modular synthesis. Eurorack is characterized by its compact size, 3.5mm mono jacks, and cables for patching signals. Control Voltage (CV) is a fundamental aspect of Eurorack systems, allowing users to modulate parameters like pitch, filter cutoff, and amplitude using sources such as LFOs, envelopes, gates, and more.

It was decided early on to use cloud-based APIs for the speech-to-text, text-to-speech and general knowledge response functions because it significantly reduces CPU usage on the local device, freeing up resources for audio processing, reducing latency and improving response quality. An audio effect was created as an experiment to test various software and hardware platforms to meet certain criteria; the effect should be able to change the sound of a voice to a new voice, with a convincing result, and be able to run in multiple instances to accommodate the simulation of a choir of voices. The basis for the effect is formant change paired with subtle changes in pitch and timing. The effect uses techniques of Fast Fourier Transformation (FFT) computing and sampling for the processing. In the selection process, the voice changer effect

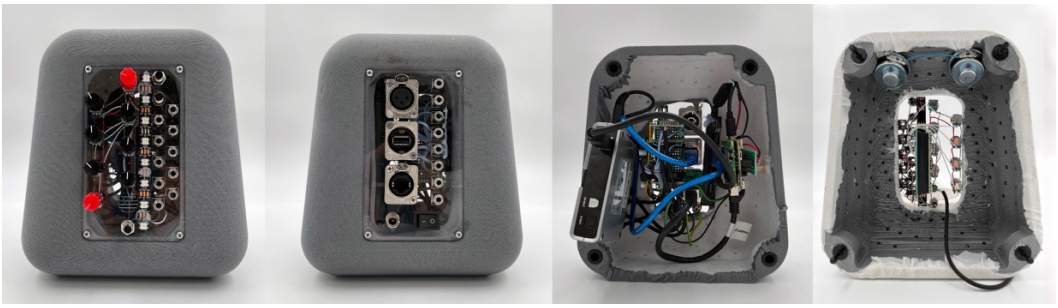

Figure 2: Left: The final version of MoSS (front and back). Right: Inside the case.

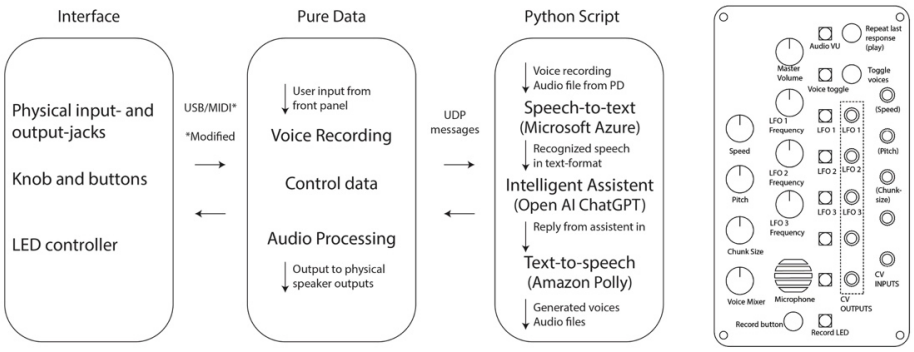

Figure 3: Left; Overview of the different technical components in MoSS. Right: overview of the user interface.

was created on microcontrollers, which would be useful as a post-processing effect on hacking of existing hardware. Teensy 4.1, Daisy Seed, ESP32 were all candidates for the experiment. However, it was also found that the system should be self-contained and thus be able to run multiple services at once. For this reason, the system of choice became a single board computer with custom software to control the smart speaker functions and another set of functions for the sound processing. The voice changing effect experiment was then carried out on a Raspberry Pi using Miller Puckette's Pure Data, and the FFTease library.

4 THE MOSS PROTOTYPE: TECHNICAL DESCRIPTION

In this section we will go more into detail with the final version and technical setup of MoSS. As shown in Fig. 2, MoSS borrows certain design aesthetics of a typical smart speaker: a rounded symmetrical shape with fabric covering. The polished design ideal of a typical smart speaker is, however, broken by the panels on the front and back of the device. The clear acrylic exposes the electronic guts of the unit, with clear references to hacking and tinkering. The knobs, buttons and jack connectors on the front panel derive from the concept of pairing a smart speaker with a modular synthesizer. The layout allows for easy, hands-on control of its functions. Minijack patch cables can be used to connect its in- and outputs to sensors, LFO-modules etc. The interface also provides visual feedback through a series of addressable LEDs. Inside the unit, a Raspberry Pi along with a custom-built interface controller and an audio interface is the base of the system. In addition, there are amplified speaker outputs on the back of the device for connecting up to eight separate speakers. Further, it is possible to also connect different kinds of sensors that can provide data which also may affect the interaction. The case is composed of two identical shells, held together with friction fit pegs and sockets that come apart with little force, making it simple to access the insides without the requirement of tools.

The MoSS prototype uses the Microsoft Azure API for speech-to-text, the OpenAI ChatGPT API as intelligent assistant and Amazon Polly API for text-to-speech. The audio recording as well as processing is done with Pure Data. Everything is tied together with a Python script. (see Fig. 3). In tests, Microsoft Azure performed well in both accuracy and speed for Danish language. The results were both faster and more accurate when compared to Google's STT service. ChatGPT was chosen for its straight-forward API, where practically all customizations can be done with direct text prompting. This adds the extra possibility of changing the settings directly in the Python script running on MoSS. Amazon Polly was chosen mainly because of the extensive audio parameters in SSML tags. On the software side the device is configured with 64-bit Raspbian as the operating system. Simultaneously to the voice assistant software, the device runs a program that handles recording and processing audio and communication with the hardware interface. The smart speaker software is also modular in its structure in that it is possible to change or replace a process with another API or built-in software.

To activate MoSS, the user presses a button to start talking to the intelligent assistant, which after a few moments will respond with an answer. There are currently three Danish voices available from the Amazon Polly API, and the script will generate a separate audio file for each of the three. Once the files are downloaded, they are played back in Pure Data, with real-time control of the running audio processing patch. Playback for each voice is done using looped sampling, with two simultaneously looping "playback heads" that are 180 degrees out of phase. This sampling allows for real-time time compression, expansion, and transposition. Looped sampling also introduces a variety of sonic artifacts that allow for a wide range of sound manipulation with simple controls. It is controlled by three buttons on the user interface. One for pitch, one for speed, and one for "chunk size"; the duration of the transition between the looping playhead (Fig. 3, right).

5 THE MOSS WORKSHOP

We invited six experts in interaction design, music and programming, to participate in a full-day workshop in November 2023 to test different aspects of MoSS. First, we presented the interface of the MoSS prototype; the different features, knobs, buttons, and possibilities for connecting sensors and speakers through the physical input and output jacks in the prototype. Then we asked participants to engage with the different possibilities for modifying the vocal expressions of the speech synthesis smart speaker system, in terms of e.g. pitch, volume and speed. To end the first part of the workshop, we discussed in plenum the participants' experiences of the playability and possibilities of exploring the aesthetic aspects of the vocal expressions of the device by turning knobs, connecting sensors and using a variety of different speakers. We started the second part of the workshop presenting how to work with the audio manipulation and filtering in Pure Data and Python, which the participants could use to playfully investigate and design specific vocal expressions. The participants used Pure Data for audio manipulation and filtering of e.g. the speed, pitch and timbre and audio effects such as reverb and delay. In Python the participants could manipulate text prompts for ChatGPT and Amazon Polly, e.g. insert SSML tags, changing input and output text and the behavior of the general knowledge response generation. To wrap up the second part of the workshop, we discussed in plenum the participants' experiences of the possibilities of building and customizing the aesthetics of the vocal expressions by changing the code, and when and how this backend customization of the smart speaker synthesizer device could help create a meaningful, playful and imaginative interaction to them.

During the workshop we observed the participants' investigations of the MoSS prototype. We also interviewed all the participants to get deeper reflections on what meaningful and playful interactions they had in relation to the MoSS prototype as interlocutors, designers and programmers. Here, we were particularly interested in what it meant for the participant's experience of the smart speaker device to be able to play and modify the modular speech synthesis.

6 FINDINGS AND CONCLUDING REMARKS

Through our observations during the workshop and the interviews with participants, we gained some preliminary findings. First, some of the participants found that the MoSS prototype invited playful experiments through manipulating a singular parameter, e.g. the stretching of the timing of the synthetic voice. It was more difficult for participants to know what was happening when they modified several parameters at the same time. Second, the playful investigations sparked reflections among the participants on different vocal imaginaries triggered by their explorations of the possible sonic landscapes for the synthetic voices e.g. a playful exploration of time-stretching a synthetic voice led to reflections on how time is experienced differently by different species (e.g. trees, birds, humans). By time-stretching the synthetic voice, some participants started to discuss if they could imagine vocal expressions of the synthetic voice channeling the time of trees or birds and not just human time. This also led to considerations of how synthetic voices have the potential to evoke different vocal identities that might also be associated to the more-than-human, making the experience of interacting with the smart speaker more imaginative and reflective. Third, this way of engaging with the smart speaker opened for aesthetic curiosities and speculations, ultimately changing the smart speaker from solely a functional technology to also becoming a playable instrument for vocal expressions. Rather than being passive recipients of pre-set vocal expressions as part of the smart speaker design, the participants could interact and explore the vocal expressions and voice qualities as interactive and playable elements of the smart speaker device, making it more engaging, and fun.

In the workshop, we saw how MoSS enabled experimentation and imagination of novel forms of vocality in speech synthesis systems. In a future iteration, we are planning to deploy MoSS in a probe study, where people can interact with the device over time. We will explore how this playful engagement with a potentially broader spectrum of vocal expressions might contribute to opening vocal imaginaries that operate beyond normative vocality towards a more pluralistic vision of synthetic voice design.

ACKNOWLEDGMENTS

Thanks to all workshop participants and the Affective Interactions & Relations (AIR) Lab at ITU. The project was funded by the Independent Research Fund Denmark under the grant ID 10.46540/2027-00236B.

REFERENCES

[1] Alice Baird, Stina Hasse Jørgensen, Emilia Parada-Cabaleiro, Nicholas Cummings, Simone Hantke, and Björn Schüller. 2018. The Perception of Vocal Traits in Synthesized Voices: Age, Gender, and Human Likeness. *Journal of the Audio Engineering Society* 66, 4: 277–285.

[2] Julia Cambre and Chinmay Kulkarni. 2019. One Voice Fits All?: Social Implications and Research Challenges of Designing Voices for Smart Devices. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW: 1–19.

[3] K. R. Chowdhary. 2020. Introducing Artificial Intelligence. In *Fundamentals of Artificial Intelligence*. Springer India, New Delhi, 1–23. Chelsea Finn. 2018. Learning to Learn with Gradients. PhD Thesis, EECS Department, University of Berkeley.

[4] Leigh Clark, Philip Doyle, Diego Garaialde, *et al.* 2019. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers* 31, 4: 349–371.Matthew Van Gundy, Davide Balzarotti, and Giovanni Vigna. 2007. Catch me, if you can: Evading network signatures with web-based polymorphic worms. In Proceedings of the first USENIX workshop on Offensive Technologies (WOOT '07) . USENIX Association, Berkley, CA, Article 7, 9 pages.

[5] Andreea Danielescu, Sharone A Horowit-Hendler, Alexandria Pabst, Kenneth Michael Stewart, Eric M Gallo, and Matthew Peter Aylett. 2023. Creating Inclusive Voices for the 21st Century: A Non-Binary Text-to-Speech for Conversational Assistants. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ACM, 1–17.

[6] Inchan Jung, Hankyung Kim, and Youn-kyung Lim. 2021. Understanding How Users Experience the Physiological Expression of Non-humanoid Voice-based Conversational Agent in Healthcare Services. *Designing Interactive Systems Conference 2021*, ACM, 1433–1446.

[7] Edward B Kang. 2022. Biometric imaginaries: Formatting voice, body, identity to data. *Social Studies of Science* 52, 4: 581–602.

[8] Christine Murad, Cosmin Munteanu, Benjamin R. Cowan, and Leigh Clark. 2021. Finding a New Voice: Transitioning Designers from GUI to VUI Design. *CUI 2021 - 3rd Conference on Conversational User Interfaces*, ACM, 1–12.

[9] Thao Phan. 2017. The Materiality of the Digital and the Gendered Voice of Siri. *Transformations Journal* 29: 11. 10.

[10] Miguel Sicart. 2022. Playthings. *Games and Culture* 17, 1: 140–155.