to avoid the stereotypical monotonous robot-like speech, there is little attention being paid to the consequences that result from the current skeuomorphic synthesized vocal identities being created through these methods.

There is a need to study and discuss the politics of the digitally designed voices in the era of algorithmic computation, especially in relation to what representation and identities are evoked through the increasing prevalence of synthesized voices in everyday life, as well as to explore alternative representational strategies to the existing synthesized voices.

Here we will present our project, which is driven by the question; Since machines are not limited to a single identity, why do machine voices have only one gender, one age, and one accent? As an alternative to the existing skeuomorphic synthesized voices, we seek to develop a collective synthesized voice enabling participation in corpus design, and engaging different everyday people in the development process in terms of open-source frameworks. Here it is important to point out that we do not believe that one participant necessarily equals one gender, but that this participant can represent multiple genders as we do not conceive of gender as something essentialist or static, but as fluid and temporally unfolded (Jones 2012).

In the following a short presentation of voice synthesis technology will be given, followed by an introduction to the feminist and queer theories of gender that have motivated our research and design project. Then we present our project, [multi'vocal], and our methodological approach with a specific focus on the design choices and principles implemented in order to create a collective synthesized voice, by enabling everyday people to engage in the corpus collection and open-source frameworks for a gender non-binary collective synthesized voice. Our project will then briefly be contextualized though an introduction to related work in the field. Finally, a short discussion of the challenges that emerge when designing for creating a collective non-binary synthesized voice are made, followed by a conclusion and a reflection on future work possibilities.

2. LITERATURE REVIEW: NON-BINARY GENDER AND SYNTHESIZED VOICES

2.1 Synthesized voices

Text-to-speech (TTS) systems are a common way to generate synthesized speech, allowing users to freely input their desired text. Such systems are often integrating adaptations of Time-Domain Pitch Synchronous OverLap- Add (TD-PSOA) based, concatenative synthesis (Dutoit & Leich 1993), a method which requires a large corpus of pre-recorded human speech, segmented into phoneme sized instances, allocated from a text input, forming synthesized replications of the sentences (Boughazi & Tabet 2011). Parametric speech synthesis, also used in TTS systems, has the addition of DNN's, to computationally improve the selection process, allowing for more human-like features (Barra-Chicote et al. 2010). Additionally, some DNN based frameworks, such as WaveNet (Zen et al. 2016), are now challenging the state-of-the-art for speech synthesis.

With corpus collection being the driver for such systems, a crucial element for engineers during this stage is to achieve consistency across the entire corpus, both for pronunciation and phrasing (Saratxaga et al. 2006). This is important in order to achieve intelligibility, as segments may be taken from differing instances, depending on the text input (FestVox 2014a). In order to achieve such qualities recordings are made in laboratory conditions or studios, with recordings for a corpus size with 3000+ sentences (FestVox 2014b). These digitally designed voices are often categorized in terms of language and the gender represented (Baird et al. 2017).

2.2 Non-binary Gender

Following the philosopher and queer theorist Judith Butler the gender binary is part of a regulatory practice that upholds the heterosexual hegemony through acts that reiterate the binary of male and female. These normative constraints produce a regulation of who comes to matter as socially visible subjects (Butler 1993).

As it is argued in Baird et al. (2017) and Phan (2017) synthesized voices such as Apple's Siri or the 13 different voices in the IBM Watson archive are categorised within a binary system as either male or female. The politics of digitally designed voices in the era of algorithmic computation lies among other things, in the way in which these voices are categorized within a gender binary system. Because this can be argued to demonstrate a normative understanding of gender within a strict binary frame, where the synthesized voices are rubricated in a classification system that divides gender into two distinct categories of male and female, regarded as oppositional and disconnected. In a study by Baird et al. (2018) however, the experience of gender in synthesized voices are shown not to be merely a binary decision by listeners. This is an important observation that challenges the state-of-the-art for how researchers and companies classify and communicate the identity constructions of synthesized voices.

In the following we will present how theories of the non-binary gender identifications have inspired our development of the project, [multi'vocal], then we describe how specific focus have guided our design choices and the principles implemented in order to engage everyday people in the corpus collection and open-source frameworks for gender non-binary collective synthesized voice.

3. DESCRIPTION OF PROJECT: [MULTI'VOCAL]

In this section we present our project, [multi'vocal], and how it has been inspired by queer feminist thinking. Then we describe how the project has been designed in relation to the technical specifications required when creating an open-source synthesized voice. Here we specifically present our methodological approach in the quest to create a non-binary gender synthesized voice by inviting people with all genders, ages and accents to contribute their voice to the corpus collection. We describe how this strategy is supported through our design and principles of how to create the technical system. From a technical perspective our design, research and art collective (henceforth mv) aims to create a technology enabling an infinitely expanding synthesized voice, composed from an increasing corpus of diverse vocal identities. This section provides an introduction to how queer feminist thinking have inspired our project [multi'vocal], the design and technical framework, and finally to related work.

3.1. [multi'vocal]: non-binary synthesized voice

Our project, [multi'vocal], can be defined as a feminist activist project that explores how synthesized voices can be named, categorised and produced differently. In our project we seek to challenge the gender binary categorisation and production of synthesized voices, by creating a non-binary voice where all genders (ages and accents) presented by the voices from participants are heard. The non-binary is a category for gender identities that are outside or not solely male or female, not conforming to the norms of the gender binary. Our aim is to push the ways in which we think, talk and frame representation and identity constructions in technologies such as synthesized voices. As an alternative to the existing synthesized voices, we want to create a synthesized voice with a proliferation and multiplication of gender categories, not static but as fluid and temporally unfolding categories. Through this we want to explore if it is possible to create a representational technology that enhances identification as process, complicating fixed binary identities; a non-binary voice that moves beyond the fixed binary identity.

Identification as process is a term discussed by cultural theorist Stuart Hall, who wrote "Perhaps instead of thinking of identity as an already accomplished fact, which the new cultural practices then represent, we should think, instead, of identity as a 'production' which is never complete, always in process, and always constituted within, not outside, representation" (1994 p. 392). Art history scholar Amelia Jones has explained identification as process as "how subjects might navigate the world through process rather than endless oppositional projections that seek to fix others in place in order to confirm the self" (2012 p. 229).

With our project, [multi'vocal] we aim to create a synthesized voice that enhance identification as process, by inviting many different people with all genders, accents and ages to contribute their voice to the corpus collection and design of the synthesized voice. This we hope will complicate the fixed binary identities that are currently dominating the field of voice synthesis.

3.2 Design and technical frameworks

mv has prototyped a design of a data acquisition recording system (henceforth: [multi'vocal] system), which deviates from conventional corporate methods as presented in the 'Introduction', encouraging both participation and discussion in the field of voice synthesis, engaging the participants in the project in public places, events and on online open-source fora such as github.

The design choices and principles implemented by mv, are in order to enable the engagement and participation of everyday people in the corpus collection, and open-source frameworks for collective synthesized voices are presented in the following.

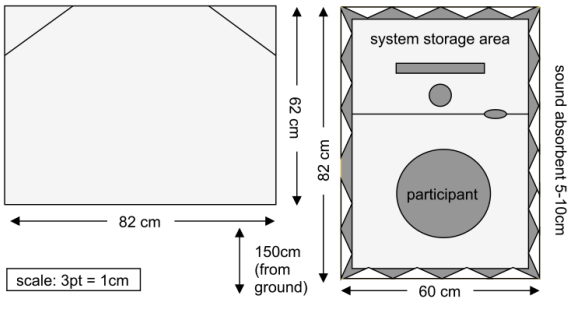

Figure 1: *The dimensions of the [multi'vocal] system, from side (left), and above (right) views. Materials and equipment were left at the bare minimum to ensure an easily replicable design, and protect against unattended conditions.*

3.2.1. Design

When developing the physical structure priorities for the [multi'vocal] system have included robustness, affordability and modularity; so that the level in which individual components of the [multi'vocal] system is possible to reconstruct for everyday people around (most of) the world. Dimension details of the [multi'vocal] system are shown in Figure 1.

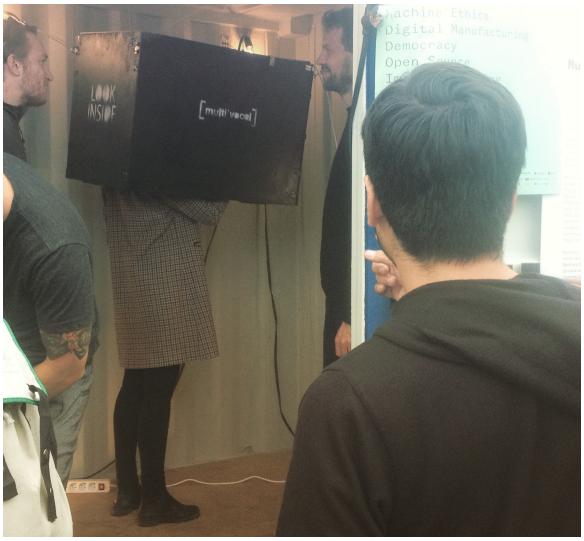

Figure 2: *The [multi'vocal] system in use. Different people from all over the world are able to enter the system whilst standing, and contribute their voice to the [multi'vocal] synthesized voice corpus collected at international festivals and events.*

The physical structure consists of a head height suspended container (Cf. Figure 2) open at the bottom to allow participants to enter. The structure is placed at height covering to at least the shoulders (blocking peripheral view), with a footstool available for differing heights. The exterior of the [multi'vocal] system is made entirely from plywood, in order to keep the overall mass lightweight, whilst retaining robustness, given the types of locations it will be positioned in. Underneath the wooden exterior 5cm of sound insulation material is fixed, in order to reduce the external background noise.

When participants enter the [multi'vocal] system the screen itself is placed on the back wall, and around average eyes level. The microphone is directly below this, and the button to the right is fixed to an acrylic glass screen, used to protect all elements.

3.2.2. Technical frameworks

For the front-end much like reasons outlined in the Section 'Design', robustness, modularity and affordability were prioritised, so off-the-shelf equipment were chosen. The full list of front-end equipment (minus cabling) is as follows: Dynamic microphone (Shure SM57),11" monitor display, 2 channel audio interface (M-Audio Fasttrack Pro), Single-board computer (Raspberry Pi), Microcontroller (Arduino) connected to Push Button.

At its core mv aims to make modularity of frameworks possible, contributing to the open-source community and providing full documentation in each stage. Both front and back-end code is freely available through the mv github.

Figure 3: *Simplistic, Graphical User Interface design, implemented inside the recording system. Showing an example sentence read aloud by participants. A full data set of phonetically and prosodically balanced sentence transcriptions is available from (FestVox 2014c).*

At run-time the GUI shows a white screen with black and grey text (Cf. Figure 3), a minimalist choice to avoid overloading the view of the participant. Simplicity in the design and experience is aimed at in order to invite as many different participants as possible (Feinberg & Murphy 2000). When the participant holds the record push button, the recording begins and on release the recording ends. The resulting wav file (sample rate of 44.1kHz, and bit rate of 16), is then added to an offline uploading queue and the next sentence transcription is immediately loaded to the display. Since Internet was not always secure, this approach ensured that recordings were not lost.

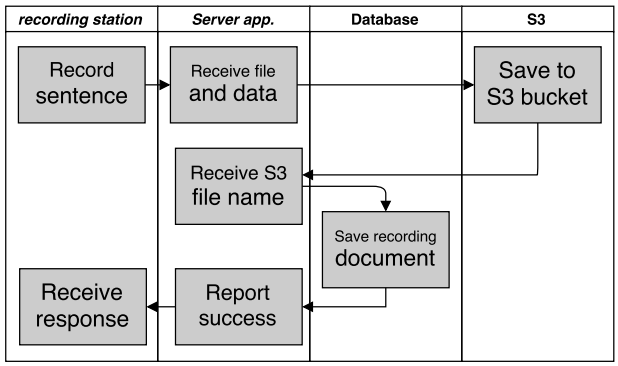

Figure 4: *[multi'vocal] system back-end processing, from upload to storage, applied to each recorded speech instances.*

The [multi'vocal] system runs all back-end (Cf. Figure 4.) components on Amazon Web Services(AWS) (Amazon 2017a); recordings are saved directly onto a Simple Storage Service (S3) (Amazon 2017b) bucket, and the recording application is running on an Elastic Cloud Computer (EC2) (Amazon 2017c) instance. Mass storage in S3 is inexpensive, enabling mv to indiscriminately save all received recordings.

Verification and management of the continually increasing corpus runs on one of the smallest possible instances on AWS, allowing for the [multi'vocal] system to run inexpensively. When the application receives a recording, the file is saved to the s3 bucket, inserting a document, which includes the name of the file and metadata. The metadata consists of;

- client identifier (ID) - a unique ID for that recording instance, as well as a location ID, so that data collected from differing events or sources can later be managed.
- transcription ID - to indicate the transcription related to the recorded instance.

No other metadata is collected on the individual recordings. However, since all contain the client ID data this still allows mv to differentiate between recordings to produce varying synthesized voice corpus compositions.

3.3 Related Work

Data collection, specifically for speech informs many Interactive Systems, including TTS and Automatic Speech Recognition (ASR). Predominately online, there are projects now exploring different avenues of this. Although to the best of the authors knowledge there are no projects working on diverse corpus construction, specifically for the synthesized voice, aside from [multi'vocal] and the Mozilla Corporation, a non-profit open source technology giant, that is now working on the project Common Voice (Mozilla Corporation 2017) an online collaborative data collection system which gathers speech recordings from diverse people across the world to contribute to an open source ASR system. With speech recognition being limited to the corpus it is trained on, this project is the other side of the coin to [multi'vocal], aiming to improve diversity specifically in ASR, removing the limitation of current state-of-the-art corpora.

4. ENGAGEMENT OF EVERYDAY PEOPLE IN THE [MULTI'VOCAL] VOICE CORPUS COLLECTION

Here we will briefly discuss our experiences so far on how to engage and include the diversity of everyday people on festivals and public events in the production of the collective gender non-binary synthesized voice, [multi'vocal].

The engagement of everyday people in the development of a design are used commonly today within the field of HCI (Huybrechts 2014), but not in speech synthesis. Yet consistently voice synthesis corpora are created with blackboxed technologies and the voice corpus are gathered in controlled environments (FestVox 2014b) with one professional voice actor. In this way the development of state-of-the-art synthesized voices offers limited agency to everyday people in terms of technological transparency and identity adaptation - something which in regards to e.g. gender identification, may show to have societal consequences (Phan 2017).

The [multi'vocal] project aims to open up the field of voice synthesis to more diverse participation, enabling engagement from everyday people in the open-source frameworks hosted online, and through housing the [multi'vocal] system in environments outside of laboratory conditions, to create a collaborative synthesized voice by engaging everyday people from all over the world. So far utilising public festivals and events such as Roskilde Festival 2017, Techfestival 2017, and Catch at CLICK festival 2018 (now it the project is installed at Catch an art and innovation space, where it will stay until end of 2018), for speech corpora collection, has insured an inclusion of individuals with diverse backgrounds, across a short period of time. Typically, the recordings for the speech corpora are made in a controlled environment, with somewhat strict direction from engineers. mv's approach is more spontaneous (Ward 1989), where the Graphical User Interface (GUI) design (Cf. Figure 2) provides the only guidance for the person recording their voice as part of the voice corpus collection to the collective synthesized voice. However, the lack of strict guidance in relation to the engagement of everyday people in the voice corpus collection has sometimes resulted in unusable recordings with background noise or recordings of half sentences, which brings a workload to manually sorting the voice recordings.

As [multi'vocal] is political project we have also made several talks where the feminist activist agenda is presented with the aim of sparking debate of representation in current technologies.

An ongoing challenge is still however, how we can get beyond the notion of fixed categories when we are still using categories such as gender, age and accent in order to talk about the identity constructions and representations we want to explore through our project. Can we use the

categories that we critique and want to get beyond? Can data-driven systems that operate with annotations and archival categories create a place for identification at all? Or do we, by operating with categories such as gender, age and accent reiterate the norms and notions of fixed binaries? We have decided not to ask participants about their gender, age or accent when they contribute with their voice in order to avoid using fixed categories as little as possible. The problem of technologies operating and reiterating fixed gender binaries, however, need more work in the future.

5. CONCLUSION AND FUTURE WORK

In this paper we have presented [multi'vocal], a project exploring if the chosen methodology can engage and include the diversity of everyday people in the creation of a non-binary synthesized voice.

Here we have first presented the influence from feminist and queer theory discussing possibilities of creating synthesized voices beyond the gender binary enhancing identification as process instead of fixed identity constructions. Then we presented the methodological and technical considerations in terms of design choices and data collection principles, followed by a short discussion of the challenges that emerge when designing for the engagement of everyday people, especially in regards to the development of a collective non-binary synthesized voice.

One of the challenges in this project is how we make the exploration of the alternative synthesized voices available to diverse group of people with different backgrounds. The aim of the project is to include voices from people with different genders, ages and accents, yet so far we have only been able to engage people on-site at different festivals and events in Denmark. Through the development of an online platform we hope to engage with people internationally, yet we are aware that this platform will only be accessible to people who are able to go online excluding millions of people worldwide from participating in the project.

Moving forward the methodology of the [multi'vocal] project will be refined in terms of enabling everyday people to participate in the voice corpus collection online, so people not being physically present where the [multi'vocal] project is installed, are still able to contribute. Also refining the data collection methodologies will be an area for advancement in terms of noise reduction and sound quality in general. At a later point an online annotation protocol to the [multi'vocal] system back-end, storing information and recording quality, should be developed. Another future development will be to use the knowledge gained from making the [multi'vocal] project in the English language, to enable participation of everyday people in the development of synthesized voices in other languages as well. This goal also connect to the overall aim for the mv collective, creating a methodology that can enable engagement of everyday people in the development of a collective non-binary synthesized voices, adding to the diversity of synthesized voices in all languages.

We are also concerned with whether the everyday people participating in the project on festivals and events are aware of the aim with [multi'vocal]; that our project is a feminist activist project that should spark discussion and make participants reflect upon if and how they want identity constructions and representations in technologies to be now and in the future. We have communicated our project aims in written text making the aim explicit. This text is shown next to the voice booth and on websites of the festivals and events where the project is exhibited and presented. In the future the text will possibly stress the fact that we do not believe that one participant equals one gender, but that gender is performative and temporal and can change over time. Due to this a participant is invited to contribute as many times as they wish, expressing multiple identifications.

We hope that we will be able to make a mobile [multi'vocal] voice recording booth, so that we can engage everyday people in other countries and in places where people are not so socially or physically mobile, in order to collect voices to the [multi'vocal] speech corpora.

6. ACKNOWLEDGEMENTS

We thank all the participants who have contributed with their voices on festivals and venues, and all the mv collective friends without whom [multi'vocal] would never have been possible.

7. REFERENCES

Amazon. 2017a. Start Building on AWS Today. Amazon Web Services. Available from: https://aws.amazon.com/ (retrieved 4 June 2018).

Amazon. 2017b. Amazon S3. Amazon Web Services. Available from: https://aws.amazon.com/s3/ (retrieved 3 June 2018).

Amazon. 2017c. Amazon EC2. Amazon Web Services. Available from: https://aws.amazon.com/ec2/ (retrieved 4 June 2018).

Baird, A., Jørgensen, S.H., Parada-Cabeliero, E., Hantke, S., Cummins, N., Schuller, B. (2017) Perception of Paralinguistic Traits in Synthesized Voices. In: Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences (AM 2017), London, UK, 23-25 August 2017. New York, USA: ACM. No pagination.

Baird, A., Jørgensen, S.H., Parada-Cabeliero, E., Hantke, S., Cummins, N., Schuller, B. (2018) The Perception of Vocal Traits in Synthesized Voices: Age, Gender, and Human-Likeness. The Journal of Audio Engineering Society, Special Issue on Augmented and Participatory Sound and Music Interaction using Semantic Audio, 66 (4), 277-285.

Barra-Chicote, R., Yamagishi, J., King, S., Montero, J. M., Macias-Guarasa, J. (2010) Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. Speech Communication, 52 (5), 394–404.

Butler, J. (1993) Bodies that Matter. New York: Routledge.

Clifford, N., L. Kwan Min. (2000) Does Computer-generated Speech Manifest Personality? An Experimental Test of Similarity-attraction. In: Proceeding of the SIGCHI Conference on Human Factors in Computing Systems (CHI '00). New York, USA: ACM. 329–336.

Mozilla Corporation. 2017. Project Common Voice. Mozilla Corporation. Available from: https://voice.mozilla.org/da (retrieved 4 June 2018).

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.

Feinberg, S., Murphy M. (2000) Applying Cognitive Load Theory to the Design of Web-based Instruction. In: Proceeding of the 18th Annual ACM International Conference on Computer Documentation: Technology & Teamwork (SIGDOC '00). Cambridge, USA: IEEE. 353–360.

FestVox. 2014a. Building Synthetic Voices: Building Prosodic Models. FestVox. Available from: http://festvox.org/bsv/c1639.html (retrieved 4 June 2018).

FestVox. 2014b. Building Synthetic Voices: Corpus Development. FestVox. Available from: http://festvox.org/bsv/c2176.html (retrieved 4 June 2018).

FestVox. 2014c. CMU Arctic. FestVox. Avaialble from: http://festvox.org/cmu_arctic/cmuarctic.data (retrieved 4 June 2018).

Gagnon, R. (1978) Votrax real time hardware for phoneme synthesis of speech. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'78). Tulsa, Oklahoma, USA. Cambridge, USA: IEEE, no pagination.

Hall, S. (1994) Cultural Identity and Diaspora. In: Patrick Williams & Laura Chrisman (eds.). Cultural Identity and Diaspora. In Colonial Discourse and Post-Colonial Theory: A reader. New York: Columbia University Press, 392-404.

Hunt, A.J., Black, A.W. (1996) Unit selection in a concatenative speech synthesis system using a large speech database. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96), Atlanta, GA, USA. Cambrigde, USA: IEEE. 1520–6149.

Huybrechts L. (ed.) (2014) Participation Is Risky: Approaches to Joint Creative Processes Amsterdam: Valiz.

Iida, A., Campbell, N., Higuchi, F., Yasumura, M. (2003) A corpus-based speech synthesis system with emotion. In: Speech Communication, 40 (1-2), 161–187.

Jones, A. (2012) Seeing Differently: A history and theory of identification and the visual arts. New York: Routledge.

Phan, T. (2017) The Materiality of the Digital and the Gendered Voice of Siri. Transformations 29, 23–33.

Saratxaga, I., Navas, E., Hernáez, I., Luengo I. (2006) Designing and recording an emotional speech database for corpus based synthesis in Basque. In: Proceedings of 5th international conference on Language Resources and Evaluation (LREC'06). Genoa, Italy. Luxemburg: European Language Resources Association (ELRA). 2126–2129.

Schroeder, M. R. (1966) Vocoders: Analysis and synthesis of speech. The Bell System Technical Journal, 54 (5), 720–734.

Peterson, G. E., Wang, W., Sivertsen, E. (1958) Segmentation techniques in speech synthesis. The Journal of the Acoustical Society of America, 30, 739–742.

Tabet, Y., Boughazi, M. (2011) Speech synthesis techniques. A survey. Proceedings of 7th International Workshop on Systems, Signal Processing and their Applications (WOSSPA'11), Tipaza, Algeria. Cambridge, USA: IEEE. No pagination.

Ward, W. (1989) Understanding Spontaneous Speech. In: Proceeding of the Workshop on Speech and Natural Language (HLT '89).

Stroudsburg, PA, USA: Association for Computational Linguistics. 137–141.

Young, S. J. (1979) Speech synthesis from concept: A method for speech output from information systems. The Journal of the Acoustical Society of America, 66 (3), 685.

Zen, H., Senior, A., Schuster, M. (2013) Statistical parametric speech synthesis using deep neural networks. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'13). Vancouver, BC, Canada. Cambridge, USA: IEEE. No pagination.

Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K., (2016) Wavenet: A generative model for raw audio. arXiv, 1609:03499, Google DeepMind, London, UK.